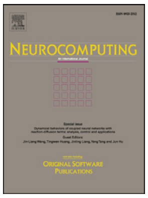
Neurocomputing 322 (2018) 166–176

Contents lists available at ScienceDirect

Neurocomputing

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



EFUI: An ensemble framework using uncertain inference for pornographic image recognition

Rongbo Shen[a,b], Fuhao Zou[a,b], Jingkuan Song[c], Kezhou Yan[d], Ke Zhou[a,b,*]

[a] *Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System, Luoyu Road 1037, Wuhan, PR China*
[b] *School of Computer Science and Technology, Huazhong University of Science and Technology of China, Luoyu Road 1037, Wuhan, PR China*
[c] *University of Electronic Science and Technology of China, No. 2006, Xiyuan Ave., Chengdu, Sichuan, China*
[d] *Tencent Inc., Shenzhen, Guangdong, PR China*

ARTICLE INFO

Article history:
Received 24 March 2018
Revised 14 August 2018
Accepted 15 August 2018
Available online 22 September 2018

Communicated by Prof. X. Gao

Keywords:
Pornographic image recognition
Ensemble framework
Uncertain inference
Bayesian network
Deep learning

ABSTRACT

Pornographic image recognition is a challenging task due to the subjective definition and complex contextual information. In this paper, we propose an ensemble framework using uncertain inference (EFUI) for pornographic image recognition. The EFUI employs bayesian network (BN) as uncertain inference engine, while prior global confidence and uncertain evidence of local semantic components are acquired by deep learning networks. Specifically, we construct the graphical model of BN based on the internal contextual relationship of local semantic components, which conforms to common sense. The prior global confidence of pornography for candidate image is extracted using GoogleNet/ResNet-50. We extract the uncertain evidence of local semantic components, represented by the probability of visual object presence, using Single Shot MultiBox Detector (SSD). Finally, a novel uncertain belief propagation algorithm is introduced to propagate the belief of extracted uncertain evidence until convergence, then identifying the pornographic possibility of the image according to the final confidence. To evaluate EFUI, we employ the public NPDI Pornography database and collect a practical dataset contains 10 million images. Experimental results on all datasets demonstrate that our EFUI achieves the state-of-the-art performance on all evaluation metrics, and outperforms the best counterparts by up 6.95% and 0.61% for accuracy on the two datasets respectively. It also presents a comparable performance over the state-of-the-art approaches in literature.

1. Introduction

Recently, the user generated contents increase rapidly and usually contain Not Suitable/Safe For Work (NSFW) [1] images, especially pornographic images, which are often prohibited in public environments and for particular population (*e.g.*, children) [2]. With the development of online media and digital mobile device, the fast growth of such contents spread into the internet easily. A report by the ExtremeTech technology website suggested that 30% of all Internet traffic was associated with adult contents and some adult websites had several times page views than large-scale news websites [3]. Because manual discrimination is tedious and inefficient, automatic recognition is the popular solution. However, the definition of pornography is relative subjective and pornography generally contains diverse semantic components with complex contextual relationship, so it is difficult to exactly recognize pornographic images.

Among various types of NSFW images, pornographic images are often the most unwelcome. A series of approaches have been proposed to detect them for over two decades. The early approach adopted a strategy to detect nudity based on the color and texture properties of human skin [4], followed by the identification of nude body parts using geometrical analysis. Then, an appropriate classifier was employed to further filter the images. This approach could lead to a lot of false positives, because some images may expose a lot of skin in activities with nothing about pornography (*e.g.*,swimming, sunbathing). The skin detector is vulnerable to materials with skin likely colors, and the geometrical analysis also suffers possible failure of body parts identification. Inspired by the bag-of-words models in text classification, an image can be represented by an unsorted set of discrete visual words, and this representation has appeared promising for image recognition [5]. More recently, many bag-of-visual-words (BoVW) approaches employed discrete local descriptors to survey images, and encoded

[*] Corresponding author at: Luoyu Road 1037, Wuhan, PR China.
E-mail addresses: rock_shen@hust.edu.cn (R. Shen), fuhao_zou@hust.edu.cn (F. Zou), jingkuan.song@gmail.com (J. Song), kezhouyan@tencent.com (K. Yan), k.zhou@hust.edu.cn (K. Zhou).

https://doi.org/10.1016/j.neucom.2018.08.080
0925-2312/© 2018 Published by Elsevier B.V.

the extracted features into intermediate representation. Then, the images were classified into different pornographic categories by a SVM classifier. These approaches still suffer some ambiguous problems, such as the definition of task-specific visual codebooks and mapping the low-level features to visual words. In recent years, deep learning has achieved great breakthroughs in various image recognition tasks (*e.g.*, [6,7]), especially Convolutional Neural Networks (CNN). Specifically, CNN combines many diverse layers for automatic features extraction and classification (*i.e.*, no need for hand-crafted features). Therefore, many approaches based on off-the-shelf CNNs were proposed for automatic pornographic image recognition. However, these approaches did not explore contextual relationship of semantic components in pornographic images. The semantic components usually include local visual objects, color and texture.

In general, each image has only one main topic and contains diverse semantic components. Some previous works [8–10] have proved that a generative graphical model [11] was suitable to describe the contextual relationship of semantic components, and could facilitate the main topic recognition. The detected semantic components by machine are usually represented as the probability of presence, that introduce uncertainty. Hence, we propose an ensemble framework using uncertain inference (EFUI) for pornographic image recognition. The framework employs bayesian network (BN) [12] as uncertain inference [13] engine, while prior global confidence and uncertain evidence of local semantic components are obtained by deep learning networks. The proposed EFUI can utilize prior global confidence and semantic components with contextual relationship jointly. Specifically, we construct the graphical model of BN based on the internal contextual relationship of local semantic components, which conforms to common sense. The prior global confidence of pornography for candidate image is extracted using GoogleNet [14]/ResNet-50 [15]. We extract uncertain evidence [16,17] of local semantic components, represented by the probability of visual object presence, using Single Shot MultiBox Detector (SSD) [18]. Finally, a novel uncertain belief propagation algorithm is introduced to propagate the belief of extracted uncertain evidence until convergence, then identifying the pornographic possibility of the image according to the final confidence.

In order to evaluate the performance of the proposed EFUI and compare with the state-of-the-art approaches, we employ the public NPDI Pornography database [19] and collect a practical dataset contains 10 million images from social internet media (*e.g.*, Tencent WeChat and QQ). Experimental results on all datasets demonstrate that our EFUI achieves the state-of-the-art performance on all evaluation metrics, and outperforms the best counterparts by up 6.95% and 0.61% for accuracy on the two datasets respectively. It also presents a comparable performance over the state-of-the-art approaches in literature.

Although this work focus on the pornographic image recognition, the methodology we discuss herein is easily extended to other type of applications, such as scene recognition. We summarize our contributions as follows:

- We propose a novel ensemble framework using uncertain inference (EFUI) for pornographic image recognition. The EFUI employs the internal contextual relationship of local semantic components to facilitate the recognition.
- We introduce a novel uncertain belief propagation algorithm to propagate the belief of extracted uncertain evidence.
- We collect a practical dataset contains 10 million images from social internet media. We annotate 9250 pornographic images of the practical dataset, given the location and category of our defined visual objects.

2. Related work

In this section, we review some representative pornographic image recognition approaches in literature, including *Skin-based approaches, BoVW-based approaches* and *CNN-based approaches*.

2.1. Skin-based approaches

The first efforts to detect pornography are associated with nudity, where the approaches tried to identify skin information of nude or scantily-clad body [4,20–22]. In general, these approaches classify each pixel as skin pixel or non-skin pixel to detect skin regions. Then, a geometric analysis strategy is employed to extract features from skin regions. Finally, a classifier is trained to classify the image as nudity or not.

Fleck and Forsyth [4,20] proposed a content-based retrieval strategy for recognizing images with naked people. This approach combined color and texture properties to obtain an effective mask for skin regions. These skin regions were then grouped to nude body parts using geometric analysis. This approach was demonstrated to have 60% precision and 52% recall on a test set of 138 uncontrolled images of naked people. Jones and Rehg [21] constructed a statistical color models for skin and non-skin classes from a large dataset of labeled pixels. A histogram of 256 bins for each color channel is computed from the skin images, and another for the non-skin. From these histograms, five different features are extracted to train a decision tree classifier. Zaidan et al. [22] proposed an anti-pornography system with two stages, namely skin detection stage and pornography recognition stage. In the first stage, a multi-agent learning method combined the bayesian method and the back-propagation neural network, to extract skin regions from the image accurately with taking into consideration of various special conditions. In the second stage, the features from the skin regions were extracted to classify the images into either pornographic or non-pornographic.

However, these approaches present relatively poor performance, due to some non-pornographic images with much exposed skin in activities (*e.g.*,swimming, sunbathing).

2.2. BoVW-based approaches

More recently, inspired by the bag-of-words models in text classification, many bag-of-visual-words (BoVW) approaches were proposed to tackle pornography recognition problems [19,23–28]. Generally, these approaches employ different feature descriptors to extract key points with features in image. Based on the training set, a visual codebook can be learned by k-means algorithm. According to the visual codebook, the features of key points are converted into an intermediate representation using uniform feature vector. Finally, a SVM classifier can be trained to classify images.

Deselaers et al. [23] firstly proposed a BoVW model using local features to classify pornography into different categories. In first step, local features were extracted around interest points by SIFT descriptors. Next, local features were encoded into an intermediate representation based on visual codebooks. Codewords were selected through gaussian mixture models, generating the visual codebooks. Finally, a SVM classifier was trained to classify the images. Lopes et al. [24] proposed a BoVW-based approach to perform image classification using Hue-SIFT descriptors, a SIFT extension which included color information. A comparison test was also made between the SIFT descriptors and Hue-SIFT descriptors applied to pornographic images, showing that the combination of color information and local features performed better. Ulges and Stahl [25] employed visual words with color-enhanced discrete cosine transform (DCT) descriptors in YUV color space for representing images. Steel [26] proposed a BoVW-based nudity detection by using a gaussian skin masking, namely Mask-SIFT descriptors. All pixels of an image without relation to skin were removed before feature transform. Based on skin, shape and local features, the author developed a cascading classifier to classify the images. Avila et al. [19] proposed an extension of the BoVW-based approach for the pornographic video detection task. Key frames were extracted from segmented shots to represent the video, then Hue-SIFT descriptors were extracted on a dense spatial grid. The BossaNova mid-level image representation was used to encode the local features, followed by a SVM classifier for each key frames and a majority voting scheme for video class prediction. Caetano et al. [27] proposed an approach for pornography detection based on local binary feature extraction and BossaNova image representation, a BoVW model extension that preserved more richly visual information. Moreira et al. [28] introduced a space-temporal interest point detector and descriptor, namely Temporal Robust Features (TRoF), to aggregate local information into a mid-level representation using Fisher Vectors, the state-of-the-art model of BoVW.

These BoVW-based approaches significantly improved the performance of the pornographic image recognition than skin-based approaches. These approaches still suffer some ambiguous problems, such as the definition of task-specific visual codebooks and mapping the low-level features to visual words.

2.3. CNN-based approaches

In recent years, many works adopt off-the-shelf CNN models for pornography recognition and produce superior performance without hand-crafted visual feature descriptors [29–33]. Generally, these approaches employ transfer learning to fine-tune an end-to-end CNN model.

Moustafa [29] proposed a combination of AlexNet [34] and GoogleNet [14] models to classify selected frames, integrating the final result for a video via majority voting scheme. The author adjusted the last layer in AlexNet and GoogleNet models, and employed transfer learning on the pre-trained models trained from ImageNet database [35]. Nian et al. [30] proposed a novel scheme utilizing CNN to detect pornographic images efficiently with a single model. This work employ two strategies to promote the efficient of training algorithm: transfer learning from pre-trained model, adjusting the training data at appropriate time according to the performance on validation set. Al-Shabi et al. [31] proposed a mixture of CNN models for adult content recognition. This work was formulated on a weighted sum of multiple CNN models, the weights of CNN models were learned using linear regression. Mahadeokar et al. [32] from Yahoo release a trained ResNet-50 [15] model for NSFW image recognition, particularly pornographic images in the wild. This model also employed transfer learning to fine-tune to fine-tune pre-trained model. Perez et al. [33] proposed a novel CNN method of combining static and motion information for video pornography detection. The motion information can alleviate the ambiguity of pornography recognition problem. The authors designed an architecture related to the two-stream CNN model, one pathway for the static information and another for the motion information. The pathways were later combined by score averaging.

The CNN-based approaches can present excellent performance. However, these approaches did not explore contextual relationship of semantic components in pornographic images, which can further promote the performance of recognition.

3. Methodology

Our proposed EFUI includes four components: *Prior global confidence extraction, Uncertain evidence extraction, Ensemble framework scheme* and *Uncertain belief propagation algorithm*. The overview of the proposed EFUI is shown in the Fig. 1.

Table 1
The configuration of GoogleNet architecture is used for the task of prior global confidence extraction. Inception layers stack various type of layers with 2 layer depth, incorporating convolution layers with kernel size 5 × 5, 3 × 3, 1 × 1 and pooling layer with kernel size 3 × 3.

Type	Kernel size & stride	Output size
Convolution	7 × 7 & 2	112 × 112 × 64
Max pool	3 × 3 & 2	56 × 56 × 64
Convolution	3 × 3 & 1	56 × 56 × 192
Max pool	3 × 3 & 2	28 × 28 × 192
Inception(3a)		28 × 28 × 256
Inception(3b)		28 × 28 × 480
Max pool	3 × 3 & 2	14 × 14 × 480
Inception(4a)		14 × 14 × 512
Inception(4b)		14 × 14 × 512
Inception(4c)		14 × 14 × 512
Inception(4d)		14 × 14 × 528
Inception(4e)		14 × 14 × 832
Max pool	3 × 3 & 2	7 × 7 × 832
Inception(5a)		7 × 7 × 832
Inception(5b)		7 × 7 × 1024
Avg pool	7 × 7 & 1	1 × 1 × 1024
Dropout(40%)		1 × 1 × 1024
Linear		1 × 1 × 2
Softmax		1 × 1 × 2

3.1. Prior global confidence

The prior global confidence is the prior probability of category that this image belong to. Two state-of-the-art deep learning networks, namely GoogleNet [14] and ResNet-50 [15], are employed to obtain the prior global confidence. GoogleNet is a kind of specific CNN model, which inception module provides excellent ability of feature representation, mapping and abstraction. ResNet-50 is a 50 layers version of residual CNN model, which residual learning blocks with shortcut connections can provide very depth of feature represents. In order to speed up the convergence of model training, we adopt the transfer learning strategy to initialize the weights of models from pre-trained models. This strategy avoids to train CNN models from scratch. In this study, the pre-trained models employ the GoogleNet trained on ImageNet [35] database and the ResNet-50 of Yahoo for NSFW image recognition.

Take GoogleNet for example, we train a reconfigured GoogleNet on pornographic image database for the task of prior global confidence extraction. Specifically, the outputs of softmax layers in the model are reconfigured to 2, and the other layers inherit the same structure from the pre-trained model. Furthermore, we initialize the weights of our model from the pre-trained model before fine-tuning. The configuration of GoogleNet architecture is shown in Table 1. The input image should be resized and cropped to the size of 224 × 224 × 3. It is notable that inception layers stack various type of layers with 2 layer depth, incorporating convolution layers with kernel size 5 × 5, 3 × 3, 1 × 1 and pooling layer with kernel size 3 × 3. Likewise, the training of ResNet-50 is adopted the same strategy.

Assuming that I is an input image, G is the GoogleNet/ResNet-50 model, and θ donates the parameter of G. The prior global confidence r of image I is a binomial distribution, given by

$$r = G(I, \theta). \quad (1)$$

3.2. Uncertain evidence extraction

The uncertain evidence of semantic components is represented by the probability of visual object presence. A state-of-the-art deep

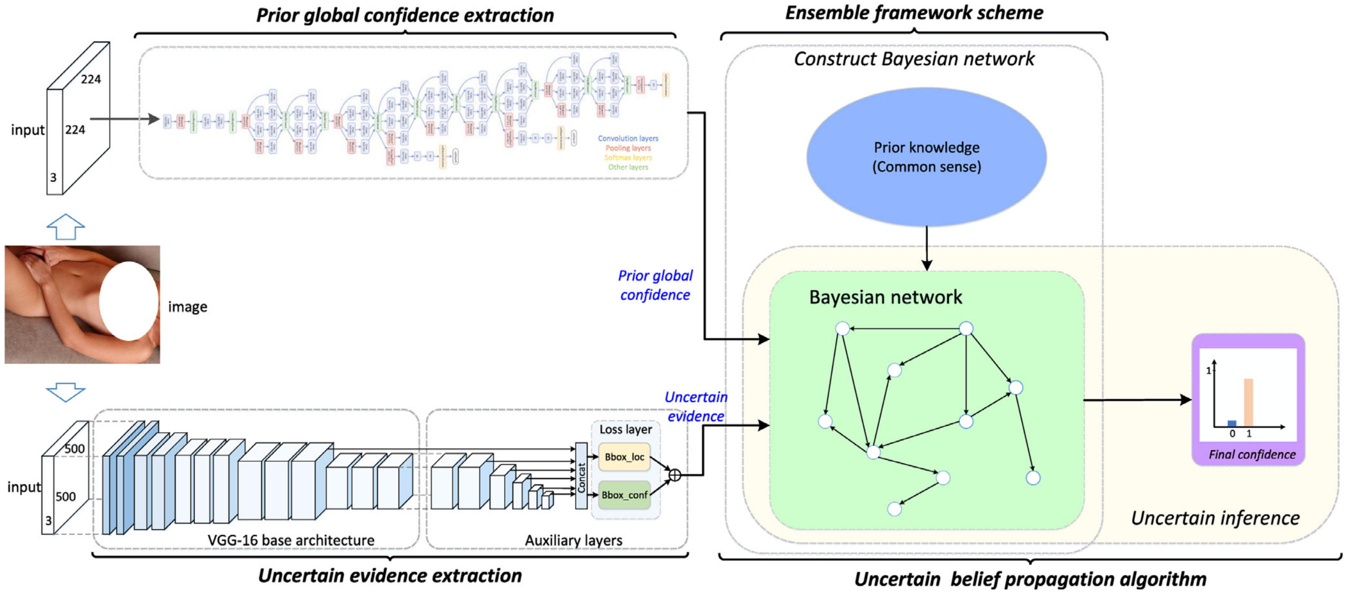


Fig. 1. The overview of the proposed EFUI. The proposed EFUI includes four components: *Prior global confidence extraction, Uncertain evidence extraction, Ensemble framework scheme* and *Uncertain belief propagation algorithm*. (1) Upper branch: The state-of-the-art deep learning networks, namely GoogleNet [14]/ResNet-50 [15], are employed to obtain the prior global confidence. We show GoogleNet as example. (2) Lower branch: A state-of-the-art deep learning network, Single Shot MultiBox Detector (SSD) [18], is introduced to detect the uncertain evidence. (3) Ensemble framework scheme: BN is employed as uncertain inference engine. We construct the BN by exploiting the internal contextual relationship of local semantic components, that conforms to common sense. (4) Uncertain belief propagation algorithm: A novel uncertain belief propagation algorithm is introduced to propagate the belief of extracted uncertain evidence until convergence, then identifying the pornographic possibility of the image according to the final confidence.

Table 2
The visual objects are corresponding to the total of 6 sensitive semantic components, which are detected by using SSD. SSD: Single Shot MultiBox Detector is a state-of-the-art deep learning network for visual object detection.

Visual objects	Abbreviation
Naked breast	NBR
Naked male genitals	NMG
Naked female genitals	NFG
Naked ass	NA
Naked body	NBO
Sexual action	SA

learning network, Single Shot MultiBox Detector (SSD) [18], is introduced to detect visual objects with probability. SSD presents the state-of-the-art performance in the field of object detection and has faster processing speed than Faster RCNN [36]. In the architecture of SSD, it adds several auxiliary layers to the end of the VGG-16 [37] base architecture. The feature maps of these layers decrease in size progressively and allow predictions of detections at multi-scale feature maps. At each feature map, a set of default bounding boxes with fixed size and several aspect ratios are applied to yield a group of proposed bounding boxes with class scores of objects. At the end of SSD, a non-maximum suppression strategy is applied to pick out the optimal bounding boxes as final detections.

Short et al. [38] shown that the definition of pornography was relative subjective. Although most of previous related works did not provide explicit definition, there is a general description of pornography in many literature: *"Any kind of material aiming at creating or enhancing sexual feelings or thoughts in the recipient and, at the same time containing explicit exposure and/or descriptions of the genitals, and clear and explicit sexual acts."* According to this description, we define a total of 6 sensitive semantic components in the task of uncertain evidence extraction. We train a SSD to detect corresponding visual objects of the 6 sensitive semantic components. These visual objects with abbreviation are shown in Table 2.

For efficiently matching ground truth bounding boxes and default bounding boxes at training procedure of SSD, we retain those default bounding boxes which overlaps to any ground truth bounding boxes are greater than 0.5. We also employ transfer learning strategy to speed up the SSD training. A pre-trained VGG-16 model is used to initialize the weights of base architecture in SSD model. In the prediction process, SSD model detects all visual objects which class scores are greater than a threshold (*e.g.*, 0.5). Many images have multiple visual objects of the same class. Intuitively, multiple detected visual objects should be assigned higher confidence to corresponding evidence. In order to simplify the evidence extraction in this situation, we adopt an approximate strategy that retains the maximum class score as uncertain evidence.

Assuming that I is an input image, D is the SSD model, and ω donates the parameter of D. The uncertain evidence of of local semantic components e is a set of binomial distributions with respect to our defined visual objects, given by

$$e = D(I, \omega). \tag{2}$$

3.3. Ensemble framework scheme

Prior global confidence and sensitive semantic components from different methods could be used for pornographic image recognition independently. According to existing experience, semantic components with contextual relationship can promote the performance of recognition. Therefore, we adopt a ensemble strategy to utilize prior global confidence and semantic components with contextual relationship jointly. Specifically, the BN is employed as uncertain inference engine, while prior global confidence and uncertain evidence of local semantic components are acquired in previous sections by deep learning networks. BN provides an useful tool for decision making which seeks advice from available information by transmitting probabilistic belief.

In details, we construct the BN by exploiting the internal contextual relationship of local semantic components, that conforms to common sense. The structure of BN is shown in Fig. 2, where

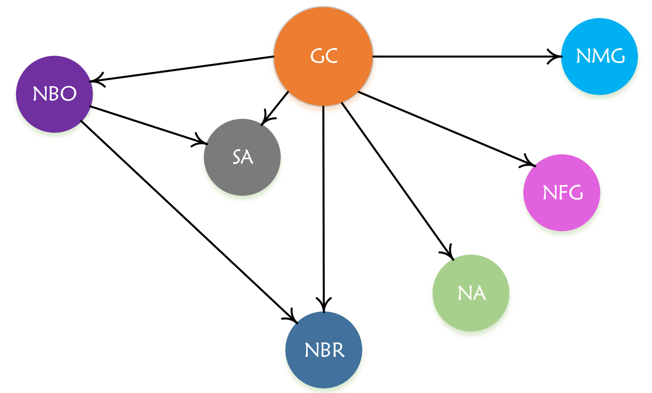


Fig. 2. The structure of BN is used for pornographic image recognition. The *GC* node denotes the global category of image, and the *NBR, NMG, NFG, NA, NBO, SA* nodes denote the state of naked breast, naked male genitals, naked female genitals, naked ass, naked body and sexual action respectively. According to the common sense, naked breast *NBR* and sexual action *SA* have strong correlation with naked body *NBO*, which have a higher probability of being accompanied by naked body.

the *GC* node denotes the global category of image, and the *NBR, NMG, NFG, NA, NBO, SA* nodes denote the state of naked breast, naked male genitals, naked female genitals, naked ass, naked body and sexual action respectively. The global category (GC) is the more global attribute of the image that contains all local semantic information, such as NBO, NA. According to the common sense, naked breast *NBR* and sexual action *SA* have strong correlation with naked body *NBO*, which have a higher probability of being accompanied by naked body. Theoretically, naked male genitals *NMG* and naked female genitals *NFG* have strong correlation with sexual action *SA*, but they have a slight probability of being accompanied with sexual action in practical image (*e.g.*, be covered). Therefore, we suppose that *NMG* and *NFG* are independent of *SA*.

After the structure of BN was determined, we should initialize the parameters of BN before inference. The parameters refer to conditional probability tables of nodes in BN. A statistical learning method is employed to calculate the initial state of conditional probability tables, which is described in the experiment setup section. In the inference process, the prior global confidence is firstly loaded to the *GC* node. Then, uncertain evidence of local semantic components is loaded to the corresponding nodes. Finally, a novel uncertain belief propagation algorithm is introduced to propagate the belief of extracted uncertain evidence until convergence.

3.4. Uncertain belief propagation algorithm

In this section, we introduce the uncertain belief propagation algorithm [39]. Belief propagation is a message passing method for inference on graphical models, such as BN. There are two type of uncertain evidence: *soft evidence* and *virtual evidence* [40–42]. The term *soft evidence* refers to evidence specified by local probability distributions that define constraints on the posterior probability distributions of one or more variables. The term *virtual evidence* is specified as a likelihood ratio, that often represents the unreliability of the evidence (*i.e.*, the observation is uncertain due to an unreliable source of information).

There are three main methods for belief propagation of uncertain evidence in BN: virtual evidence method [13], Jeffrey's Rule [43], and iterative proportional fitting procedure (IPFP) [44]. According to the evidence extraction in our proposed EFUI, the evidence is specified by explicit local probability distributions of observed variables. It should be soft evidence or hard evidence, where hard evidence refers to a specific probability distribution that one state is 1 and other states are 0. As demonstrated by Pearl [13], soft evidence can be easily converted to virtual evidence when it is on a single variable. We introduce a novel uncertain belief propagation algorithm in BN which employs the idea of IPFP and the virtual evidence method. This algorithm is an iterative procedure, and only soft evidence of one variable is considered in each iteration. The soft evidence of each variable and virtual evidence method are circularly utilized until all posterior probability distributions satisfying the evidence constraints.

Consider a BN N over a set of variables X in a particular domain, where variables are mathematical representation of nodes in BN. A joint distribution of X is defined as $P(X)$. The variables of root nodes in N are represented as R, $R \subset X$. In the structure of our BN, only one root node represents the category of image, namely main topic. H is a set of variables on which hard evidence is observed, $H \subset X$. Similarly, Y represents the set of variables on which soft evidence is observed, $Y \subset X$. Giving $Q(Y)$ as a probability distribution of variables Y.

In the first step, the prior global confidence is used to initialize the probability distribution $P(R)$ of the root node R. Because R is the only root node, which doesn't have parent nodes. The $P(X)$ can be represented by

$$P(X) = P(R)\prod\nolimits_{X_i \in X, X_i \neq R} P(X_i|\eta_i), \tag{3}$$

where X_i represents the set of variables except root node in N, η_i represents the parent nodes of X_i. Obvious, the initialization of root node need not propagate belief. Suppose h is the set of observed hard evidence of uncertain evidence e, y is the set of observed soft evidence of uncertain evidence e.

In the second step, a belief update is implemented using hard evidence h, the expression is shown as

$$P(Z) = P(X \backslash H|h), \tag{4}$$

where Z represents the set of variables in X except H.

Suppose a set of m soft evidences $SE = (se_1, se_2, \ldots, se_m)$ denotes a set of m constraints $P(Y_1|se_1), \ldots, P(Y_m|se_m)$ for m variables. A single soft evidence can be converted to virtual evidence. Given a soft evidence se_i on variable A, the virtual evidence with likelihood ratio $L(A)$ can be represented by

$$L(A) = \frac{P(A|se_i)}{P(A)} = \left(\frac{P(a_1|se_i)}{P(a_1)}, \ldots, \frac{P(a_n|se_i)}{P(a_n)}\right), \tag{5}$$

where a_i is state of A. The virtual evidence can be used to propagate belief in BN by Pearl's virtual evidence method.

In the third step, an iterative procedure is applied to circularly utilize each single soft evidence as described above. Suppose k represents the iterations, so the $j = 1 + (k-1) \bmod m$ soft evidence is used in this iteration. In the $l = 1 + \lfloor (k-1)/m \rfloor$ epoch, the virtual evidence with likelihood ratio $L_{j,\,l}(Y^j)$ can be represented by

$$\begin{aligned} L_{j,l}(Y^j) &= \frac{Q_k(Y^j|se_j)}{Q_{k-1}(Y^j)} \\ &= \left(\frac{Q_k(y^j_{(1)}|se_j)}{Q_{k-1}(y^j_{(1)})} : \frac{Q_k(y^j_{(2)}|se_j)}{Q_{k-1}(y^j_{(2)})} : \ldots : \frac{Q_k(y^j_{(s)}|se_j)}{Q_{k-1}(y^j_{(s)})}\right). \end{aligned} \tag{6}$$

Then, the update of this virtual evidence in BN can be described by

$$Q_k(Z) = Q_{k-1}(Z) \cdot \frac{Q_k(Y^j|se_j)}{Q_{k-1}(Y^j)}, \tag{7}$$

where $Q_0(Z) = P(Z)$ is the initial state.

Finally, I-divergence (also known as Kullback–Leibler distance) [45] is used as convergence termination condition for two joint distributions Q_k and Q_{k-1} over Z

$$I(Q_k||Q_{k-1}) = \sum_{z_i \in Z} Q_k(z_i) \log \frac{Q_k(z_i)}{Q_{k-1}(z_i)}. \tag{8}$$

The BN returns the posterior probability distribution of root node $Q(R|h, y)$ as the final result. The uncertain belief propagation algorithm is summarized in Algorithm 1.

Algorithm 1 uncertain belief propagation algorithm.

```
Input: The prior global confidence $r$; the set of hard evidence $h$; the set of soft evidence $y$;
Output: The posterior probability distribution of root node $R$, $Q(R|h,y)$;
Initial BN and the probability distribution $P(R)$ according to $r$;$P(X) = P(R)\prod_{X_i\in X, X_i\neq R} P(X_i|\eta_i)$;
Implement belief propagation for hard evidence $h$;$P(Z) = P(X\backslash H|h)$;
Initial $Q_0(Z) = P(Z), k = 1$;
repeat
    $j = 1 + (k-1) \bmod m$;
    $l = 1 + \lfloor (k-1)/m \rfloor$;
    $L_{j,l}(Y^j) = \frac{Q_k(Y^j|se_j)}{Q_{k-1}(Y^j)} = (\frac{Q(y^j_{(1)}|se_j)}{Q_{k-1}(y^j_{(1)})} : \frac{Q(y^j_{(2)}|se_j)}{Q_{k-1}(y^j_{(2)})} : \dots : \frac{Q(y^j_{(s)}|se_j)}{Q_{k-1}(y^j_{(s)})})$;
    $Q_k(Z) = Q_{k-1}(Z) \cdot \frac{Q(Y^j|se_j)}{Q_{k-1}(Y^j)}$;
    $k = k + 1$;
until $I(Q_k||Q_{k-1}) = \sum_{z_i\in Z} Q_k(z_i)\log\frac{Q_k(z_i)}{Q_{k-1}(z_i)}$ convergence;
return $Q(R|h,y)$;
```

4. Experiment

In this section, comprehensive experiments are described using the proposed EFUI for pornographic image recognition on the public NPDI Pornography database and the practical dataset. In order to demonstrate the performance of the proposed EFUI and compare with the state-of-the-art approaches, the counterparts employ 5 typical state-of-the-art approaches from Skin-based approaches, BoVW-based approaches and CNN-based approaches. Firstly, a brief description of datasets and evaluation metrics is given. Then, we present the experiments setup of the proposed EFUI and counterparts. Finally, a comparative analysis of experiment results is implemented.

4.1. Datasets and evaluation metrics

We employ two datasets in our experiments: the NPDI Pornography database [19] and a collected practical dataset.

The NPDI Pornography database contains nearly 80 hours of 400 pornographic and 400 non-pornographic videos. For the non-pornographic videos, 200 videos were randomly chosen from various websites and 200 videos were selected from textual queries (*e.g.*, wrestling, beach, swimming), the two subclass were called "easy" and "difficult" respectively. The videos are segmented into 16,727 shots, and a static key frame is selected to summarize the content of each shot. In total, 6387 key frames of pornographic videos, 6785 key frames of easy non-pornographic videos and 3555 key frames of difficult non-pornographic videos are selected, some samples are shown in Fig. 3. Because the key frames of pornographic videos have many non-pornographic images and the NPDI Pornography database does not give annotations of visual objects, we sample 2000 pornographic images, 1000 easy non-pornographic images and 1000 difficult non-pornographic images as evaluation dataset. The NPDI Pornography database is only used for evaluation in the experiments, the proposed EFUI on this database employs ResNet-50 model of Yahoo as prior global confidence extraction.

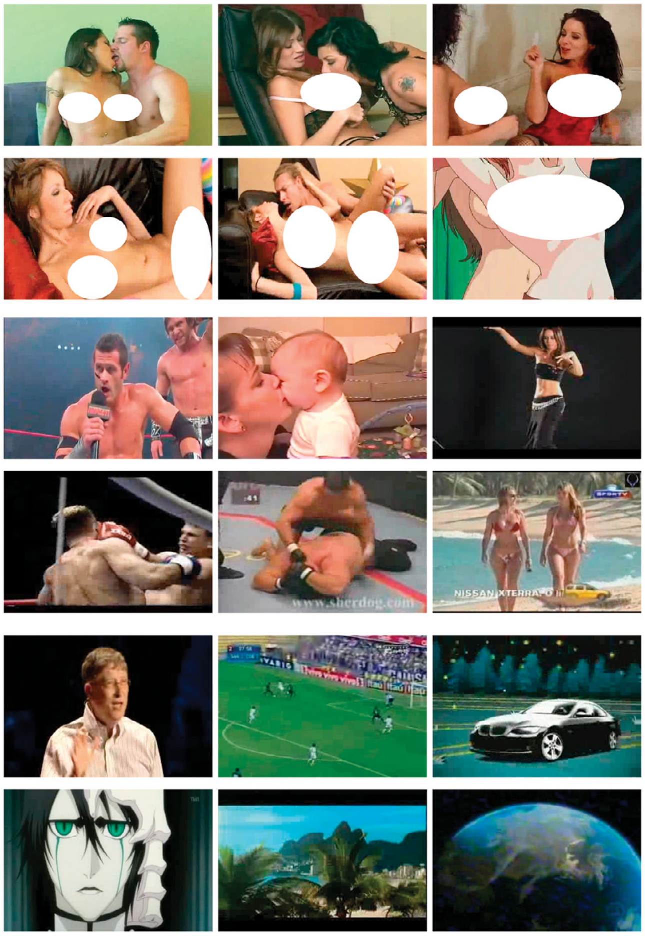


Fig. 3. Representative key frames are sampled from NPDI Pornography database. The images in the first and second rows are pornographic images, the images in the third and fourth rows are difficult non-pornographic images, the images in the last two rows are easy non-pornographic images.

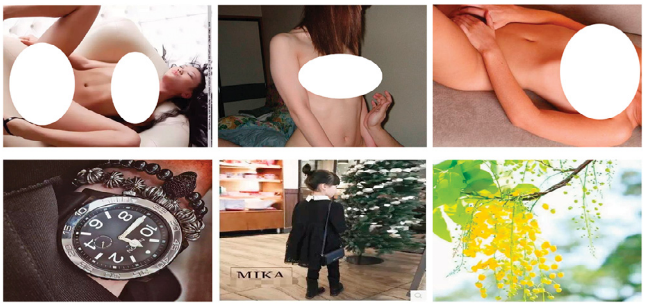


Fig. 4. Representative images from practical dataset. The images in the first row are pornographic images, the images in the second row are non-pornographic images.

In order to verify the proposed EFUI in practical engineering, we collect a practical dataset contains 10 million images from social internet media (*e.g.*, Tencent WeChat and QQ). More specifically, we acquire 5 million pornographic images and 5 million non-pornographic images by mechanical retrieval with manual review, some samples are shown in Fig. 4. We randomly choose 806,678 images from the practical dataset as evaluation dataset, which contains 399,629 non-pornographic images and 407,049 pornographic images. The rest of the practical dataset is used as training dataset. We annotate a subset of the training dataset with 9250 pornographic images, given the location and category of our defined visual objects. This subset is used for the SSD training, that is the uncertain evidence extractor in our proposed EFUI.

The task of pornographic image recognition is a two-class problem. The most common evaluation metrics related to this problem are accuracy, sensitivity (also known as true positive rate, TPR), specificity (also known as true negative rate, TNR), precision and F_1-measure [46,47]. Accuracy indicates the proportion of correct classification on both classes, and the error rate is the proportion of misclassification on both classes. Sensitivity and specificity are used to monitor the classification performance on each individual class, sensitivity for positive class and specificity for negative class. While precision is used in problems interested on highly performance on only positive class, F_1-measure is used when the performance on both classes needed to be high. F_1-measure is a comprehensive evaluation index which weighs precision and sensitivity. Beside the common evaluation metrics, we introduce the ROC curve (receiver operating characteristic curve) and AUC (area under the ROC curve) for classifiers with output of score.

4.2. Experiments setup

4.2.1. The proposed EFUI

In the proposed EFUI, the *Caffe* [48] deep learning framework is employed for training and testing of GoogleNet/ResNet-50 model and SSD model. The *Pomegranate* [49] graphical model framework is employed for uncertain inference scheme. The experimental machine has 2 NVIDIA GRID GPUs, each GPU has 2 × 2496 CUDA stream processor cores and 2 × 12 GB of video memory.

The parameter settings of learning in the proposed EFUI is a crucial factor of experiments setup. The transfer learning of GoogleNet model was fine-tuned with the weights of pre-trained model from ImageNet. The parameters of architecture are shown in Table 1. The training process had 1 million iterations of forward and backward propagation. The batch size of input layer was 256 in each iteration, and the dropout layers with 0.4 ratio were retained from the pre-trained model. The learning rate was adopted a exponential decay policy from a base value 10^{-4}, and it multiplied by decay factor 0.1 in every 0.2 million iterations. The weight decay parameter of l_2 regularization was set as 2×10^{-4}, it was employed to suppress overfitting.

The transfer learning of ResNet-50 model was fine-tuned with the weights of the ResNet-50 of Yahoo for NSFW image recognition. The training process also had 1.2 million iterations of forward and backward propagation, and the batch size of input layer was 50 in each iteration. The learning rate was adopted a exponential decay policy from a base value 10^{-4}, and it multiplied by decay factor 0.2 in every 0.2 million iterations. The weight decay parameter of l_2 regularization was also set as 2×10^{-4}.

The training of our SSD model employed a pre-trained VGG-16 model to initialize the weights of base architecture in SSD model. The total iterations of the training was 60 thousands, and the batch size was 32 in each iteration. The learning rate was also adopted a exponential decay policy from a base value 4×10^{-4}, and it multiplied by decay factor 0.1 at the 40 thousandth iteration. The weight decay parameter of l_2 regularization was set as 5×10^{-4} . The image size was set as 500 × 500, the input images were resized in the data layer. The minimum size of detected bounding boxes was set as 110 pixels, and the aspect ratio was set to a value between 0.5 and 2. It is notable that the SSD can obtain general performance in different database. Therefore, we adopt the SSD, that is trained on the subset of 9250 annotated pornographic images, as the general uncertain evidence extractor in our proposed EFUI.

The learning of BN contains structure learning and parameter learning. As describe in ensemble framework scheme of method section, the structure of BN was specified by the internal contextual relationship, it did not need to learn. The parameters of BN refer to the conditional probability distribution of each node with its parents. The joint probability distribution of BN is represented by the product of all nodes' conditional probability distribution. In our task, the state of each node is discrete and binary, its conditional probability distribution can be estimated from a training dataset by statistical learning. Theoretically, the non-pornographic images should have no visual objects, which were defined in Table 2. In practical, the SSD model has a slight possibility of false detection. In order to incorporate the prior global confidence for this situation, we prepare the training dataset with slight noisy data. We randomly choose 2000 non-pornographic images and 2000 pornographic images from the training dataset of the practical dataset, and manually annotate the visual objects. For convenience, the 2000 pornographic images can be sampled from the training dataset of SSD model, which had been annotated. The BN with parameters is shown in Fig. 5.

4.2.2. The counterparts

For the experiments on the practical dataset, the counterparts employs 5 approaches: a skin-based method using skin detector [50], a BoVW-based method using Hue-SIFT descriptors [51], a GoogleNet based CNN model, the ResNet-50 model of Yahoo [32] and the fine-tuned ResNet-50 model using our practical dataset.

For the experiments on the NPDI Pornography database, the counterparts employs 3 approaches: a skin-based method using skin detector, a BoVW-based method using Hue-SIFT descriptors and the ResNet-50 model of Yahoo. In this set of experiments, we do not employ the fine-tuned Yahoo ResNet-50 and GoogleNet using the practical dataset. There are following considerations: The practical dataset and NPDI Pornography database have different data distribution, the NPDI Pornography database has relatively more difficult samples. It is more suitable to employ original Yahoo ResNet-50 as counterpart.

The skin detector scans every pixel in the input image, and classifies it as skin pixel or non-skin pixel. Then, the skin pixels are grouped into multiple isolated regions. Finally, a geometric decision strategy is employed to analyse regions, classifying the image as nudity or not. The result of this skin detector is a binary value: true or false.

The BoVW-based method employs Hue-SIFT descriptors to extract key points with features. According to the visual codebook, the features of key points are converted into an uniform feature vector. The visual codebook can be learned from training images by k-means algorithm, the codebook size is set as 1024. Based on the uniform feature vectors from training images, a SVM classifier can be trained to classify images. For convenience, the training images adopt the training dataset of parameter learning in BN. The result of this BoVW-based approach is also a binary value: true or false.

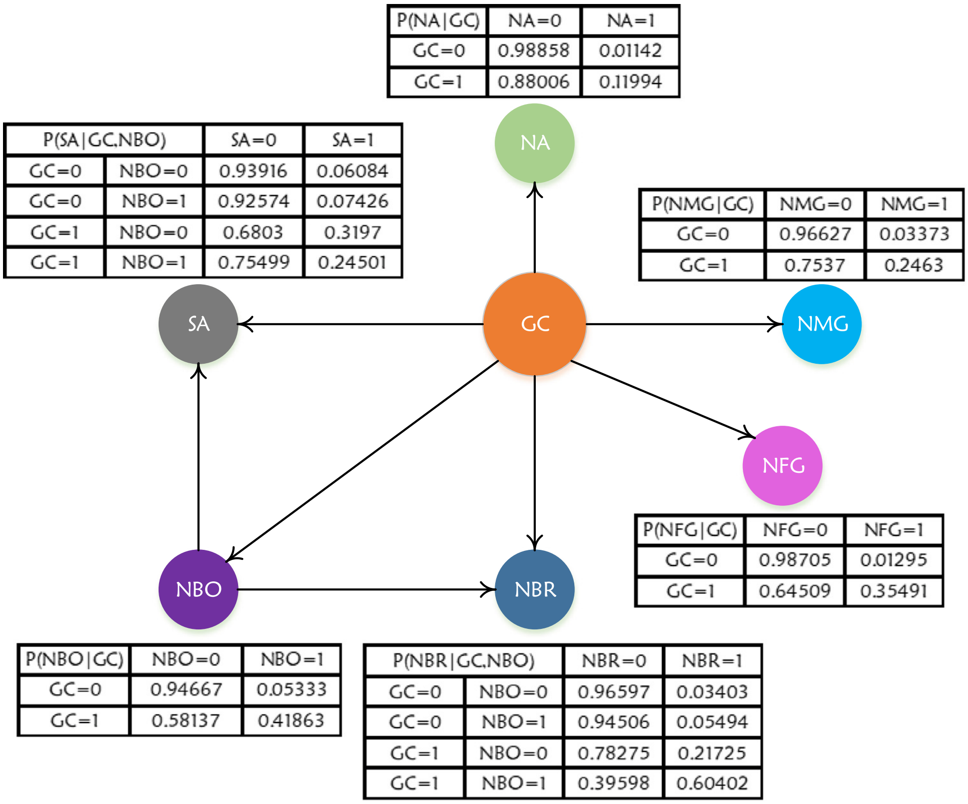
The ResNet-50 based CNN model of Yahoo can be directly used in *Caffe* [48] deep learning framework without parameter adjustment. The output is a probability between 0 to 1, the default threshold for classification is 0.5.

For a more direct comparison, we employ the deep learning networks for prior global confidence extraction. The GoogleNet based CNN model and fine-tuned ResNet-50 model are introduced in previous section.

4.3. Results and discussions

In this section, we present and discuss the results from the experiments of the proposed EFUI and the counterparts. The experimental datasets are stated at previous Datasets section.

In Tables 3 and 4, we show the evaluation performance of common metrics for each approach on practical dataset and NPDI



P(NA|GC)	NA=0	NA=1
GC=0	0.98858	0.01142
GC=1	0.88006	0.11994

P(SA|GC,NBO)		SA=0	SA=1
GC=0	NBO=0	0.93916	0.06084
GC=0	NBO=1	0.92574	0.07426
GC=1	NBO=0	0.6803	0.3197
GC=1	NBO=1	0.75499	0.24501

P(NMG|GC)	NMG=0	NMG=1
GC=0	0.96627	0.03373
GC=1	0.7537	0.2463

P(NFG|GC)	NFG=0	NFG=1
GC=0	0.98705	0.01295
GC=1	0.64509	0.35491

P(NBO|GC)	NBO=0	NBO=1
GC=0	0.94667	0.05333
GC=1	0.58137	0.41863

P(NBR|GC,NBO)		NBR=0	NBR=1
GC=0	NBO=0	0.96597	0.03403
GC=0	NBO=1	0.94506	0.05494
GC=1	NBO=0	0.78275	0.21725
GC=1	NBO=1	0.39598	0.60402

Fig. 5. The BN with parameters. The parameters of BN refer to the conditional probability distribution of each node with its parent nodes. The state of each node is discrete and binary, its conditional probability distribution can be estimated from a training dataset by statistical learning.

Table 3
The experiment results on practical dataset. The above 5 approaches belong to the counterparts. The following proposed EFUI employ GoogleNet and fine-tuned ResNet-50 as prior global confidence extraction respectively. The BoVW-based method employs Hue-SIFT descriptors and SVM classifier. The classification threshold is set as 0.5 in CNN-based approach and the proposed EFUI.

Approaches	Accuracy (%)	TPR (%)	TNR (%)	Precision (%)	F_1 (%)
Skin detector	63.27	70.27	56.14	62.00	65.88
BoVW	89.12	91.29	86.91	87.66	89.44
Yahoo ResNet-50	96.29	93.91	98.71	98.67	96.23
GoogleNet	99.13	99.36	98.90	98.92	99.14
fine-tuned ResNet-50	98.88	99.41	98.34	98.39	98.90
proposed EFUI with GoogleNet	**99.57**	**99.61**	**99.53**	**99.54**	**99.57**
proposed EFUI with ResNet-50	**99.49**	**99.65**	**99.33**	**99.34**	**99.49**

Table 4
The experiment results on NPDI Pornography database. The proposed EFUI on this database employs ResNet-50 based CNN model of Yahoo as prior global confidence extraction. The classification threshold is set as 0.5 in CNN-based method and the proposed EFUI.

Approaches	Accuracy (%)	TPR (%)	TNR (%)	Precision (%)	F_1 (%)
Skin detector	59.43	62.25	56.60	58.92	60.54
BoVW	73.10	88.80	57.40	67.58	76.75
Yahoo ResNet-50	87.75	91.60	83.90	85.05	88.20
proposed EFUI Yahoo ResNet-50	**94.70**	**95.05**	**94.35**	**94.39**	**94.72**

Pornography database respectively. The classification threshold is set as default value of 0.5 in CNN-based approaches and the proposed EFUI.

Specifically, the experiment results on practical dataset are shown in Table 3. The proposed EFUI with GoogleNet obtains 99.57% accuracy, 99.61% sensitivity (TPR), 99.53% specificity (TNR), 99.54% precision and 99.57% F_1-measure. The proposed EFUI with fine-tuned ResNet-50 obtains 99.49% accuracy, 99.65% sensitivity (TPR), 99.33% specificity (TNR), 99.34% precision and 99.49% F_1-measure. Obviously, the accuracy of proposed EFUI significantly outperforms the existing Skin-based method and BoVW-based method in counterparts, by improving approximate 36 percentage points than Skin-based approach and approximate 10 percentage points than BoVW-based approach. The GoogleNet and ResNet-50 of counterparts also present significant performance improvements than Skin-based method and BoVW-based method. These results suggest that deep learning networks present fundamental improvements in pornographic image recognition. In other side, the images in practical dataset are collected from social internet media and most of the images are well labeled, so the performance of

Table 5
The accuracy performance of the proposed EFUI and counterparts on difficult images and easy images.

Approaches	Easy images	Difficult images
Skin detector	63.80%	49.40%
BoVW	58.90%	55.90%
Yahoo ResNet-50	95.70%	72.10%
proposed EFUI Yahoo ResNet-50	**99.10%**	**89.60%**

GoogleNet and ResNet-50 tends to saturation due to the fewer difficult examples in practical dataset. Further, the proposed EFUI outperforms the GoogleNet and fine-tuned ResNet-50 in counterparts, by improving 0.44% and 0.61% respectively. Meanwhile, the decline proportions of error rates reach 50.6% than GoogleNet and 54.5% than ResNet-50 respectively. This suggests that the proposed EFUI significantly reduces the misclassification of difficult examples than GoogleNet and ResNet-50. In summary, the experiment results prove that the proposed EFUI can further improve the performance with the internal contextual relationship of local semantic components.

The experiment results on NPDI Pornography database are shown in Table 4. The proposed EFUI with Yahoo ResNet-50 obtains 94.70% accuracy, 95.05% sensitivity (TPR), 94.35% specificity (TNR), 94.39% precision and 94.72% F_1-measure. The performance of proposed EFUI significantly outperforms the existing Skin-based method and BoVW-based method in counterparts. Because the Yahoo ResNet-50 is a trained model that does not be fine-tuned on the NPDI Pornography database, its performance has slightly distinction than expectation. The Yahoo ResNet-50 still presents significant performance improvements than Skin-based method and BoVW-based method. The proposed EFUI outperforms 6.95% accuracy than the Yahoo ResNet-50 in counterparts, and the improvement of specificity (TNR) is more than 10%. This experiment demonstrates the advancement of the proposed EFUI again.

There are 2000 non-pornographic images in the experiments on NPDI Pornography database, and half of them are difficult images. Table 5 shows the accuracy performance of the proposed EFUI and counterparts on difficult images and easy images. It can be found that difficult images have higher false positive rate than easy images, because the difficult images generally have higher similarity with pornographic images and easily lead to misclassification. The results show that our proposed EFUI presents the best performance on both difficult images and easy images. Especially in the recognition of difficult images, the accuracy of our proposed EFUI significantly outperforms the counterparts, by improving approximate 17 percentage points than the ResNet-50 CNN-based model of Yahoo. This suggests that the internal contextual relationship of local semantic components can effectively improve the performance of recognition on difficult images.

The results of CNN-based approaches and the proposed EFUI can be represented by continuous probability values. The ROC curve and AUC are introduced to compare performance. The ROC curves and AUC of experiment results on practical dataset and NPDI Pornography database are shown in Figs. 6 and 7. The proposed EFUI with GoogleNet or fine-tuned ResNet-50 achieves 99.97% AUC and 99.96% AUC on practical dataset, by improving 0.02% and 0.04% than GoogleNet and fine-tuned ResNet-50 respectively. The proposed EFUI achieves 98.48% AUC on NPDI Pornography database, by improving 2.59% than the ResNet-50 CNN-based model of Yahoo.

We also review the performance of representative approaches in literature as a reference, which are shown in Table 6 with databases and solutions. Except the NPDI Pornography database, the other databases are not public. Because the key frames of pornographic videos in the NPDI Pornography database have many

Table 6
The comparison of the representative approaches in literature. It should be noted that the key frames of pornographic videos in NPDI Pornography database have many non-pornographic images, the approaches employ a partial sample frames on this database and the sampling methods are unknown.

Authors	Solutions	Databases	Performance
Fleck et al. [4]	Skin-based; geometrical analysis; threshold scheme	138 uncontrolled images; 1401 assorted control images	60% precision; 52% sensitivity
Forsyth and Fleck [20]	Skin-based; geometric constraints; threshold scheme	565 uncontrolled images; 4289 assorted control images	57% precision; 43% sensitivity
Jones and Rehg [21]	Skin-based; statistical color histogram; decision tree classifier	5241 adult images; 13970 non-adult images	80% sensitivity; 91.5% specificity
Zaidan et al. [22]	Skin-based; bayesian method; neural network	150 pornographic images; 150 non-pornographic images	96% sensitivity; 97.33% specificity
Lopes et al. [24]	BoVW-based; SIFT and Hue-SIFT descriptors; voting scheme	180 images	64.8% accuracy (SIFT); 84.6% accuracy (Hue-SIFT)
Ulges and Stahl [25]	BoVW-based; DCT descriptors; SVM	Sampled 10000 images from 5 data source	Error rate 11%-24%
Steel [26]	BoVW-based; Mask-SIFT descriptors; cascading classifier	500 pornographic images; 500 non-pornographic images	77% sensitivity,80% specificity for Mask-SIFT; 87% sensitivity,80% specificity for Cascading Classifier
Avila et al. [19]	BoVW-based; Hue-SIFT descriptors; BossaNova and SVM; voting scheme	NPDI Pornography database	96.4% accuracy of frames; 89.5% accuracy of videos
Caetano et al. [27]	BoVW-based; binary descriptors; BossaNova and SVM; multiple aggregation	NPDI Pornography database	92.4% accuracy of videos (BoW-VD); 92.0% accuracy of videos (BNVD)
Moreira et al. [28]	BoVW-based; TRoF descriptors; Fisher Vectors and SVM	NPDI Pornography database	95.18% sensitivity of videos; 95.98% specificity of videos; 95.58% accuracy of videos
Moustafa [29]	CNN-based; transfer learning; voting scheme	NPDI Pornography database	91% sensitivity,90% specificity of frames 94.1% accuracy of videos
Nian et al. [30]	CNN-based; transfer learning; novel training strategy	Unknown	98.6% accuracy
Al-Shabi et al. [31]	CNN-based; models ensemble; linear regression	6132 pornographic images; 6064 non-pornographic images	96.59% accuracy
Perez et al. [33]	CNN-based; two-stream model; optical flow and MPEG motion vectors	NPDI Pornography database	96.4% accuracy,96.7% F_2 of 2k videos; 97.9% accuracy of 800 videos

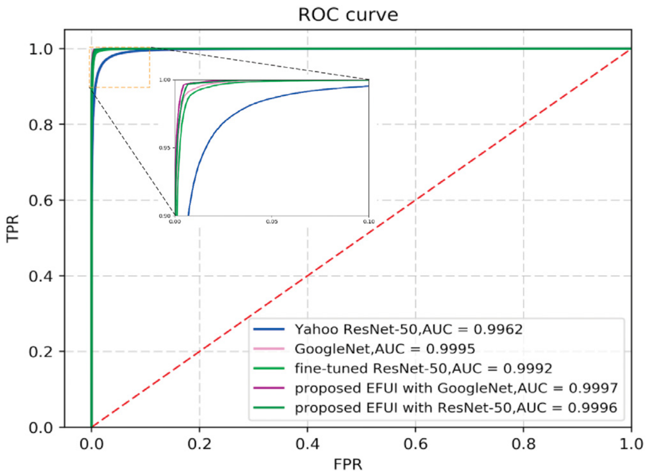


Fig. 6. The ROC curves and AUC of CNN-based approaches and our proposed EFUI are demonstrated on practical dataset. The subplot in this figure demonstrates the partial ROC curves in the FPR range [0.0,0.1] and the TPR range [0.9,1.0].

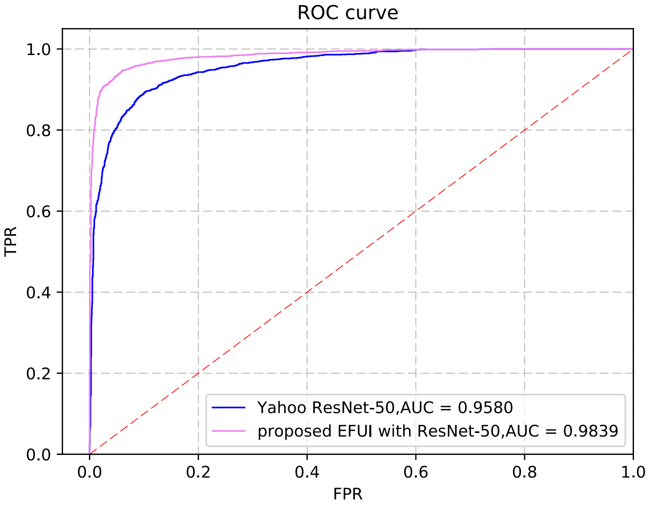


Fig. 7. The ROC curves and AUC of CNN-based method and our proposed EFUI are demonstrated on NPDI Pornography database.

non-pornographic images, the approaches in Table 6 employ a partial key frames on this database and the sampling methods are unknown. Therefore, we can not compare with these approaches directly due to the different evaluation dataset. Referencing the results in Table 6, the Yahoo ResNet-50 method in our experiments achieves similar performance with the method of Moustafa [29] on the NPDI Pornography database. Meanwhile, the proposed EFUI outperforms the Yahoo ResNet-50. Therefore, we can infer that our proposed EFUI produces a comparable performance over the state-of-the-art approaches in literature.

5. Conclusions

In this work, we propose an ensemble framework using uncertain inference for pornographic image recognition. The ensemble framework employs bayesian network as uncertain inference engine, while prior global confidence and uncertain evidence of local semantic components are acquired by deep learning networks. The experimental results on all datasets demonstrate that the EFUI can present a state-of-the-art performance for pornographic image recognition. This proves that the internal contextual relationship of local semantic components can promote the performance of recognition. We expect to make use of more richer semantic information for image recognition in the future work.

Besides applying the EFUI method in pornographic image recognition, it also can be introduced to complex scene recognition. The local semantic components are also beneficial in many other applications, such as robot vision, digital image libraries, intelligent recommendation, etc.

Acknowledgment

This work is partially supported by the Natural Science Foundation of China (No. 61672254).

References

[1] F. Attwood, I.Q. Hunter, Not Safe for Work? Teaching and Researching the Sexually Explicit Introduction, Sexualities 12 (5) (2009) 547–557.
[2] P.Y. Lee, S.C. Hui, A.C.M. Fong, An intelligent categorization engine for bilingual web content filtering, IEEE Trans. Multimed. 7 (6) (2005) 1183–1190.
[3] S. Anthony, Just how big are porn sites?, 2012, URL http://www.extremetech.com/computing/123929-just-how-big-are-porn-sites.
[4] M. Fleck, D. Forsyth, C. Bregler, Finding naked people, in: Proceedings of the Computer Vision ECCV (1996) 593–602.
[5] Y.-G. Jiang, J. Yang, C.-W. Ngo, A.G. Hauptmann, Representations of keypoint-based semantic concept detection: a comprehensive study, IEEE Trans. Multimed. 12 (1) (2010) 42–53.
[6] T. Li, B. Cheng, B. Ni, G. Liu, S. Yan, Multitask low-rank affinity graph for image segmentation and image annotation, ACM Trans. Intell. Syst. Technol. TIST 7 (4) (2016) 65.
[7] Y. Liu, J. Song, K. Zhou, L. Yan, L. Liu, F. Zou, L. Shao, Deep self-taught hashing for image retrieval, in: Proceedings of the IEEE Transactions on Cybernetics(2018).
[8] L.-J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: classification, annotation and segmentation in an automatic framework, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2036–2043.
[9] L. Fei-Fei, L.-J. Li, What, where and who? telling the story of an image by activity classification, scene recognition and object categorization, in: Proceedings of the Computer Vision, Springer, 2010, pp. 157–171.
[10] T. Li, Z. Meng, B. Ni, J. Shen, M. Wang, Robust geometric p-norm feature pooling for image classification and action recognition, Image Vis. Comput. 55 (2016) 64–76.
[11] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT press, 2009.
[12] C.M. Bishop, Pattern recognition, Mach. Learn. 128 (2006) 1–58.
[13] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 2014.
[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[16] R. Pan, Y. Peng, Z. Ding, Belief update in bayesian networks using uncertain evidence, in: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, IEEE, 2006, pp. 441–444.
[17] H. Chan, A. Darwiche, On the revision of probabilistic beliefs using uncertain evidence, Artif. Intell. 163 (1) (2005) 67–90.
[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[19] S. Avila, N. Thome, M. Cord, E. Valle, A.D.A. AraúJo, Pooling in image representation: the visual codeword point of view, Comput. Vis. Image Underst. 117 (5) (2013) 453–465.
[20] D.A. Forsyth, M.M. Fleck, Automatic detection of human nudes, Int. J. Comput. Vis. 32 (1) (1999) 63–77.
[21] M.J. Jones, J.M. Rehg, Statistical color models with application to skin detection, Int. J. Comput. Vis. 46 (1) (2002) 81–96.
[22] A.A. Zaidan, N.N. Ahmad, H.A. Karim, M. Larbani, B.B. Zaidan, A. Sali, On the multi-agent learning neural and bayesian methods in skin detector and pornography classifier: an automated anti-pornography system, Neurocomputing 131 (2014) 397–418.
[23] T. Deselaers, L. Pimenidis, H. Ney, Bag-of-visual-words models for adult image classification and filtering, in: Proceedings of the 19th International Conference on Pattern Recognition, IEEE, 2008, pp. 1–4.
[24] A.P. Lopes, S.E. de Avila, A.N. Peixoto, R.S. Oliveira, A.d.A. Araújo, A bag-of-features approach based on hue-sift descriptor for nude detection, in: Proceedings of the 17th European Signal Processing Conference, IEEE, 2009, pp. 1552–1556.
[25] A. Ulges, A. Stahl, Automatic detection of child pornography using color visual words, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2011, pp. 1–6.

[26] C.M. Steel, The mask-sift cascading classifier for pornography detection, in: Proceedings of the World Congress on Internet Security (WorldCIS), IEEE, 2012, pp. 139–142.

[27] C. Caetano, S. Avila, W.R. Schwartz, S.J.F. Guimarães, A.d.A. Araújo, A mid-level video representation based on binary descriptors: a case study for pornography detection, Neurocomputing 213 (2016) 102–114.

[28] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, Pornography classification: the hidden clues in video space–time, Forensic Sci. Int. 268 (2016) 46–61.

[29] M. Moustafa, Applying deep learning to classify pornographic images and videos, arXiv:1511.08899 (2015).

[30] F. Nian, T. Li, Y. Wang, M. Xu, J. Wu, Pornographic image detection utilizing deep convolutional neural networks, Neurocomputing 210 (2016) 283–293.

[31] M. Al-Shabi, T. Connie, A.B.J. Teoh, Adult content recognition from images using a mixture of convolutional neural networks, arXiv:1612.09506 (2016).

[32] J. Mahadeokar, S. Farfade, A. Kamat, A. Kappeler, Open NSFW model, 2016, URL https://github.com/yahoo/open_nsfw.

[33] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, Video pornography detection through deep learning techniques and motion information, Neurocomputing 230 (2017) 279–293.

[34] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[36] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).

[38] M.B. Short, L. Black, A.H. Smith, C.T. Wetterneck, D.E. Wells, A review of internet pornography use research: methodology and content from the past 10 years, Cyberpsychol. Behav. Soc. Netw. 15 (1) (2012) 13–23.

[39] Y. Peng, S. Zhang, R. Pan, Bayesian network reasoning with uncertain evidences, Int. J. Uncertain. Fuzziness Knowl. Based Syst. 18 (05) (2010) 539–564.

[40] J. Bilmes, On virtual evidence and soft evidence in bayesian networks (2004). doi: https://doi.org/10.1.1.148.1030

[41] A.B. Mrad, V. Delcroix, M.A. Maalej, S. Piechowiak, M. Abid, Uncertain evidence in bayesian networks: presentation and comparison on a simple example, in: Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2012, pp. 39–48.

[42] A.B. Mrad, V. Delcroix, S. Piechowiak, P. Leicester, M. Abid, An explication of uncertain evidence in bayesian networks: likelihood evidence and probabilistic evidence, Appl. Intell. 43 (4) (2015) 802–824.

[43] J. Pearl, Jeffreys rule, passage of experience, and neo-bayesianism, in: Knowledge Representation and Defeasible Reasoning, Springer, 1990, pp. 245–265.

[44] J. Brossard, C. Leuridan, Iterated proportional fitting procedure and infinite products of stochastic matrices, arXiv:1606.09126 (2016).

[45] I. Csiszár, I-divergence geometry of probability distributions and minimization problems, Ann. Probab. (1975) 146–158.

[46] D.M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, J. Mach. Learn. Tech. 2 (1) (2011) 37–63.

[47] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (4) (2009) 427–437.

[48] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.

[49] J. Schreiber, Pomegranate, 2016, URL https://github.com/jmschrei/pomegranate.

[50] H. Hattori, Nudity detection with python, 2017, URL https://github.com/hhatto/nude.py.

[51] K.E. van de Sande, T. Gevers, C.G. Snoek, Empowering visual categorization with the GPU, IEEE Trans. Multimed. 13 (1) (2011) 60–70.

Rongbo Shen is working toward the Ph.D. degree in Big Data and Machine Learning at Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, P.R. China. He is currently an assistant researcher with Architecture Platform Department, Tencent Inc. His current research interests include Deep Learning, Artificial Intelligence for Medicine, Computer Vision.

Fuhao Zou received B.E. degree in computer science from Huazhong Normal University, Wuhan, Hubei, China, in 1998. And received M.S. and Ph.D. in computer science and technology from Huazhong University of Science and Technology (HUST), Wuhan, Hubei, China, in 2003 and 2006. Currently, he is an associate professor with the school of computer science and technology, HUST. His research interests include deep learning, multimedia understanding and analysis, big data analysis. He is senior member of China Computer Federation (CCF) and member of IEEE, ACM.

Jingkuan Song is a full professor with University of Electronic Science and Technology of China (UESTC). He joined Columbia University as a Postdoctoral Research Scientist (2016–2017), and University of Trento as a Research Fellow (2014–2016). He obtained his Ph.D. degree in 2014 from The University of Queensland (UQ), Australia (advised by Prof. Heng Tao Shen). His research interest includes large-scale multimedia retrieval, image/video segmentation and image/video understanding using hashing, graph learning and deep learning techniques.

Kezhou Yan received the M.S. degree in Electronics and Communication Engineering from Xidian University, Xi'an, China, in 2014. He is currently a senior engineer in Tencent, Shenzhen, China. His current research interests include Artificial Intelligence for Medicine, Computer Vision, Pattern Recognition and Machine Learning.

Ke Zhou received the BE, ME, and Ph.D. degrees in computer science and technology from Huazhong University of Science and Technology (HUST), China, in 1996, 1999, and 2003, respectively. He is a professor of the School of Computer Science and Technology, HUST. His main research interests include computer architecture, cloud storage, parallel I/O and storage security. He has more than 50 publications in journals and international conferences, including Performance Evaluation, FAST, MSST, ACM MM, SYSTOR, MASCOTS and ICC. He is a member of IEEE.