Pornographic Image Recognition via Weighted Multiple Instance Learning

Xin Jin [ID], Yuhui Wang, and Xiaoyang Tan

Abstract—In the era of Internet, recognizing pornographic images is of great significance for protecting children's physical and mental health. However, this task is very challenging as the key pornographic contents (e.g., breast and private part) in an image often lie in local regions of small size. In this paper, we model each image as a bag of regions, and follow a multiple instance learning (MIL) approach to train a generic region-based recognition model. Specifically, we take into account the regions' degree of pornography, and make three main contributions. First, we show that based on very few annotations of the key pornographic contents in a training image, we can generate a bag of properly sized regions, among which the potential positive regions usually contain useful contexts that can aid recognition. Second, we present a simple quantitative measure of a region's degree of pornography, which can be used to weigh the importance of different regions in a positive image. Third, we formulate the recognition task as a weighted MIL problem under the convolutional neural network framework, with a bag probability function introduced to combine the importance of different regions. Experiments on our newly collected large scale dataset demonstrate the effectiveness of the proposed method, achieving an accuracy with 97.52% true positive rate at 1% false positive rate, tested on 100K pornographic images and 100K normal images.

Index Terms—Deep learning, multiple instance learning (MIL), pornographic image recognition.

I. Introduction

THE RAPID development of Internet has brought us great convenience in daily life, while on the other hand, made

X. Jin was with Megvii Research Nanjing, Megvii Inc. (Face++), Nanjing 210000, China. He is now with the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: x.jin@nuaa.edu.cn).

Y. Wang and X. Tan are with the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: y.wang@nuaa.edu.cn; x.tan@nuaa.edu.cn).

us easily exposed on some objectionable contents, one typical example of which is the pornography [30]. Among others (e.g., videos and texts), images are arguably the most common carrier for pornography on Internet. These pornographic images, which are typically anarchic and even illegal to sell in the adult bookstores, however, can easily and widely spread on Internet through pornographic websites, social network applications, instance messaging tools, and even e-mail attachments. Children should be prevented from browsing such images without question, while grown-ups may also not want to be exposed on them, for example when at workplaces. All these have made recognizing and filtering out pornographic images a rising concern nowadays.

While considerable progress has been made in the last two decades [9], [11], [18], [24], [27], [28], [31], recognizing pornographic images is still an open problem in computer vision. There are large intraclass variations in pornographic images, due to the changes in background, scale, scenario, and human pose. In particular, the key objectionable contents (e.g., breast and private part) that distinguish a pornographic image from normal images often lie in local regions of small size, while the cluttered background (e.g., normal body, bed, room, etc.) may sometimes make up a large portion of the image. As a consequence, when sharing similar background and human pose, the pornographic images may look very similar to some nonpornographic (sexy) images. We note that to avoid copyright issues, we do not post any pornographic image in this paper.

Due to these characteristics, an effective recognition method should be able to discover the informative local regions in pornographic images. Following this principle, most of traditional algorithms are decomposed into two phrases: 1) extracting regions of interest (ROI) from complex backgrounds and 2) calculating hybrid features from extracted ROIs for recognition [17], [30], [31], [40], [45], [47]. Since pornographic contents often come with skin exposure, skin detection techniques are commonly employed for ROI extraction [30]. However, because skin detection itself is a very challenging problem, ROI-based approaches typically have very limited generalization ability. To overcome this, Kim *et al.* [21] explicitly modeled the possible pornographic contents with a set of semantic features, such as "breast," "belly," and "bottom." However, accurately detecting these semantic features is very difficult in practice, due to their small patch support and large appearance variation.

In this paper, we propose to train a generic region-based recognition model, which does not specialize in certain body

part (e.g., breast) but aims at more general and discriminative visual patterns, for example, the female's upper body with exposed breast, the lower body with exposed private part, the close-up of sexual behavior, etc. In contrast to the semantic feature approach [21], our target regions augment private body parts or sexual acts with useful context, and thus can be more reliably recognized. Furthermore, these regions are not required to contain *complete* private body part or sexual act, and some nontypical pornographic regions (e.g., a region containing only a part of breast) will also do. In addition, our target regions typically take relatively larger positions in an image, which allows us to identify a pornographic image quickly with very few evaluations.

However, training such a region-based recognition model is a nontrivial task. Standard supervised learning requires bounding box annotations for the pornographic ROI. However, unlike the well-defined semantic features [21], the annotation process of our target regions is easily influenced by subjectivity, as different people may have different understanding of useful context. Furthermore, different regions may have different degree of pornography—intuitively, a region containing a complete breast is more pornographic than the one containing a part of breast. In some extreme cases, it is even difficult to determine the label of some ambiguous regions, i.e., a region containing a very small part of breast.

To address these issues, we follow a weakly supervised multiple instance learning (MIL) approach, which relaxes the need for direct annotation of target regions using a semi-automatic bag generation strategy, and integrates the importance (the degree of pornography) of different regions in positive bags into a unified learning model. Specifically, we make three main contributions within this MIL framework.

1) We show that based on very few annotations of the key pornographic contents[1] in an image, we can generate a bag of properly sized regions, among which the potential positive regions usually combine the key pornographic contents (e.g., breast) with useful context (e.g., upper body) that can aid recognition.

2) We define a simple quantitative measure of an arbitrary region's degree of pornography, according to its overlapping ratio with the annotated key pornographic regions. This can be naturally used to weigh the importance of different regions in a positive bag, since intuitively a region containing a complete breast should contribute larger to the bag probability than that containing a small part of breast.

3) We formulate the pornographic image recognition task as a deep weighted MIL problem, with a bag probability function introduced to combine the importance of different regions. The resulting region-based classifier is very robust and accurate, and can be conveniently used for testing by transforming the test image into a bag of multiscale regions.

---

[1]In this paper, we refer to the private body parts (e.g., breast and genital) and sexual behaviors as the *key pornographic contents* in images, as they are the key factors for distinguishing a pornographic image from normal images.

To validate the effectiveness of our approach, we have collected a large scale dataset from Internet consisting of 138 000 pornographic images and 205 000 normal images. This dataset covers almost all types of pornographic images on Internet, and intentionally includes many challenging nonpornographic images (i.e., bikini). Our method produces excellent performance on this dataset—it achieves an accuracy with 97.52% true positive rate (TPR) at 1% false positive rate (FPR) and 55 FPS with GPU, tested on 100K pornographic and 100K normal images. We also show excellent performance for cross-database experiment, by testing the pretrained model (trained on our dataset) on NPDI benchmark dataset [4].

II. Relative Work

A. Pornographic Image Recognition

There is a long history of nudity detection and pornographic image recognition in computer vision and perceptual psychology. Early works mainly focused on finding naked people in images based on a human structure model [11]–[13]. However, these works require that the pornographic images have completely naked people and simple backgrounds, while pornographic images on Internet generally have wide variations in background, scale, scenario, and human pose. To recognize general pornographic images, current methods can be roughly classified as feature-based, region-based, and body part-based.

Feature-based approaches emphasize the extraction of features from the *whole* image, and popular approaches include the bag-of-feature (BoF) approach [9], [24], [32], [36], and the deep convolutional neural networks (CNNs) approach [27], [28]. The BoF approach encodes local image features (e.g., SIFT [25]) with a visual vocabulary, and hence can capture some distinct local patterns of pornographic images. But due to the use of hand-crafted features, the BoF approach usually has limited discriminative power. By contrast, the CNN-based approach can automatically learn discriminative representations from large data. However, since they directly adapt the off-the-shelf CNN architecture to model the whole pornographic images, some crucial local details (e.g., breast) are largely ignored.

Region-based approaches extract features for recognition based on the detection of ROIs in the image. Among others, skin regions are widely considered as the ROIs for pornography [20], [31], [45], [47]. To detect skin-like pixels, the input image is typically transformed from the RGB model to the RCbCr color space, and a pixel is considered to be skin-like when it satisfies a few linear constraints in RCbCr space [14]. Based on detected skin regions, multiple hand-crafted features (e.g., color, shape, and texture features) are extracted for recognition. Compared to feature-based approaches, the region-based methods are more robust against background clutter, but run the risk of inaccurate ROI detection since skin detection is a challenging problem in itself.

Body part-based approaches define several pornography-related semantic features, such as breast, belly, and bottom, and train corresponding body part detectors for these features [21], [26]. Given an image, these body part detectors

are used to scan the image, and the detection results are then arranged in a semantic feature vector for classification. However, these body part detectors are plagued by the problem of ambiguity and very likely to generate false positive detection, due to their small patch support and large appearance variations in training.

In this paper, we propose to train a generic region-based model for pornographic image recognition. Unlike traditional region-based approaches, our approach does not rely on ROI detection, and directly identify pornographic images by evaluating our model on the test image at different positions and scales. Furthermore, while similar in spirit with the body part-based approaches [21], [26], our generic recognition model aims at more general and discriminative pornographic contents in images, which typically combine useful context with the key pornographic contents and thus can be more reliably recognized than the individual private body parts.

B. Multiple Instance Learning

Research on MIL studies the problem where a real-world object is associated with a single class label but described by a number of instances. The MIL framework was first formalized by Dietterich *et al.* [10] to investigate drug activity prediction, and then have been applied to diverse applications, including image classification [5], [6], [38]; face detection [41], [42]; image annotation [16], [34]; and saliency detection [37]. Since space does not allow for a full review, we will only focus of MIL algorithms most related to ours.

MIL has been widely used for object recognition and localization, by considering each image as a "bag" of examples given by tentative object windows, and assuming that positive images contain at least one positive instance object window, while negative images only contain negative windows [7]. Traditional MIL algorithms assume that all of the instances contribute equally and independently to the bag's label. But, sometimes, it is desirable to take into account the instance importance in the learning procedure. For example, Zhang and Song [44] proposed an online weighted MIL algorithm for visual tracking. They defined a novel bag probability function that combines the weighted instance probability, leading to a more robust and much faster tracker. Our approach is also based on weighted MIL, but differs from [44] in bag generation, weight definition, and most importantly, while [44] performed an online MIL boosting technique for feature selection, we formulate the weighted MIL algorithm under the CNN framework to learn discriminative features.

To the best of our knowledge, Li *et al.* [23] is the only one to employ MIL technique for pornographic image recognition. They treat each image as a bag, and low-level visual features of divided sub-blocks as instances. Specifically, they first perform spatial pyramid partition to divide an image into blocks, and then extract three low-level visual features (i.e., color features, LBP [29], and HOG [8]) on these blocks, resulting a representation consisting of three different multi-instance bags. However, each bag in [23] is then transformed into a single representation vector (metadata), and a standard single instance learning method is used to solved the MIL problem.

By contrast, we treat each image as a bag of regions throughout the algorithm, and learn a region-based classifier that can be use not only to identify pornographic images but also to localize the rough pornographic regions in these images.

III. APPROACH

As mentioned before, the key pornographic contents in an image often lie in local regions, and the background regions may be distractive for recognition. As a consequence, recognition methods based on the whole image are very sensitive to background clutter. Considering this, we propose to model each image as a bag of regions, and follow an MIL approach to train a generic region-based recognition model.

In this section, we first describe our bag generation strategy, which is based on the annotations of the key pornographic contents in images during training. Based on these annotations, we further present a quantitative measure of the regions' degree of pornography in positive bags. With the region-based representation and pornography measure, we formulate the recognition task as a weighted MIL problem under the CNN framework.

A. Bag Generation

Bag generation is crucial for an MIL algorithm [46]. In this section, we mainly focus on the bag generation strategy in training, while the testing process will be detailed in Section III-D. Currently, the most popular approach is to consider each image as a bag of examples given by region proposals [35], [48]. However, as mentioned before, our target regions are the general pornographic regions that combine pornographic contents with useful contexts while excluding distractive backgrounds. These regions (e.g., the female's upper body with exposed breasts) typically cannot be regarded as "objects," and thus are not in accordance with the goal of region proposal algorithms.

Besides, we argue that a good positive bag in training should strive to cover various regions that may encounter in testing, including the typical pornographic regions, the nontypical positive regions and the negative (nonpornographic) regions. The nontypical positive regions, such as the upper body containing only a part of breast, are very important since we cannot always obtain typical pornographic regions for a test image. The negative regions in positive images are also very useful, since they are by nature difficult training instances. In addition, scale normalization is important for model training, and regions from different positive bags should be of similar scale. To satisfy these requirements, we propose to annotate the key pornographic contents in positive images first, and then use these annotations to guide the bag generation process.

There are two main types of the key pornographic contents, i.e., the private body parts and sexual behaviors. We perform bounding box annotation for these key pornographic contents in the training set, and for the sexual behavior, we just annotate the most compact regions that are sufficient to characterize these behaviors. In general, there are only a small number of annotations (less than ten) required for a pornographic image, and the annotation process is unlikely to be affected

by subjectivity due to the clear definition of key pornographic contents.

With these annotations, we can obtain a variety of regions for a positive image by moving a window around each annotated pornographic region. In particular, the window size is adaptively computed according to size of each annotated pornographic region, and the aspect ratio is the same to that of the image.[2] In our current system, we randomly set the window size to be 2, 2.5, or 3 times of the annotated region. Here, the slight scale change is introduced to improve model's robustness against scale variation, and these three similar scales are selected by cross validation, and thus are experimentally proved to be able to include useful context information.

By randomly moving the windows from each annotated region with different displacements, we can generate about 100 regions for a positive image, among which some are typical positive regions, some are nontypical positive regions, and some are nonpornographic regions. For negative images, the bag generation process is the same with the test procedure, as described in Section III-D.

B. Pornography Measurement

Intuitively, the typical positive regions should be more pornographic than the nontypical ones. That is, a region containing a complete breast should contribute larger to the bag probability than the one containing only a small part of breast. We take this into account, and present a simple quantitative measure of an arbitrary region's degree of pornography.

Our measurement is based on the overlapping ratio between the generated regions and the annotated key pornographic regions. We set a region's degree of pornography to be a scalar number between 0 and 1, by finding the maximum overlapping ratio between this region and the annotated key pornographic regions. Furthermore, by normalizing these degrees of pornography (i.e., sum to 1), we can obtain the weights (importance) of different regions in a pornographic image. In the following, we will describe how to integrate the importance of different regions into a deep learning framework.

C. Deep Weighted MIL

Inspired by the great success of CNN in image classification [22], [33], in this section, we formulate the recognition task as a weighted MIL problem under the CNN framework. In particular, we introduce a bag probability function to combine the importance of different region in positive images.

In our MIL setting, each image $\mathbf{X}$ is modeled as a bag of n regions $\{\mathbf{x}_i | i = 1, \ldots, n\}$. Given a region, the deep CNN extracts layer-wise representations from the first convolutional layer to the last fully connected layer. Our CNN architecture is inspired by the GoogLeNet model [33]. The output of the last fully connected layer is a 1000 dimensional vector, followed by a softmax layer to transform it into a probability distribution for objects of 1000 categories. As pornographic image recognition is a two-class classification problem, we redesign

[2]In both training and testing phases, the regions are resized to 224 × 224 to satisfy the requirement of GoogLeNet model [33].

the output of the last fully connected layer to be a 1-D vector, and transform it into a Bernoulli distribution via the sigmoid function. Formally, given region $\mathbf{x}_i$, we denote the output of the last fully connected layer as h_i, and let p_i^+ and p_i^- be the probability that region $\mathbf{x}_i$ is pornographic or nonpornographic, respectively. Then, we have

$$
\begin{aligned}
p_i^+ &= \frac{e^{h_i}}{e^{h_i} + 1} \\
p_i^- &= \frac{1}{e^{h_i} + 1}.
\end{aligned}
\tag{1}
$$

Now, let us take into account of the importance of different regions in a positive bag, and aggregate the instance probabilities (1) into a positive bag probability. As aforementioned, we can obtain the weight w_i for each region $\mathbf{x}_i$ by normalizing the degree of pornography for all extracted regions. Intuitively, we can directly assign these weights to the instance probabilities (1), and combine them to form the bag probability for a positive bag. However, we notice that the weights of some regions in a positive bag are just 0, i.e., these regions have no overlapping with the annotated key pornographic regions, and hence should be treated as negative instances. Considering this, we divide each positive bag into two sub-bags, i.e., a positive sub-bag, where the regions are all assigned with positive weights, and a negative sub-bag, where the weights are all 0. Let n^+ and n^- be the number of the regions in the positive sub-bag and negative sub-bag, respectively, we can define the sub-bag probabilities as follows:

$$
\begin{aligned}
p^+ &= \sum_{i=1}^{n^+} w_i p_i^+ \\
p^- &= \frac{1}{n^-} \sum_{i=1}^{n^-} p_i^-
\end{aligned}
\tag{2}
$$

where the probability p^- for the negative sub-bag is computed by just averaging the instance probabilities. With the sub-bag probabilities (2), we simply define the positive bag probability function as

$$
p = p^+ p^-.
\tag{3}
$$

Then, we reach the following loss function (negative bag log-likelihood function) for a positive image:

$$
L = -1\{n^+ > 0\} \log p^+ - 1\{n^- > 0\} \log p^-.
\tag{4}
$$

Actually, we notice that the bag probability function and the loss function for a negative image can also be written as (3) and (4), since we always have $n^+ = 0$ and $n^- > 0$ for a negative image. The gradient for (4) is calculated via back propagation

$$
\frac{\partial L}{\partial h_i} = \left(-1\{w_i > 0\} \frac{1}{p^+} w_i + 1\{w_i = 0\} \frac{1}{p^- n^-} \right) p_{1,i} p_{0,i}.
\tag{5}
$$

We resize each region to 224 × 224, and use the open-source package Caffe [19] to extract deep features based on the GoogLeNet model [33]. We redefine the objective of GoogLeNet as (4), and perform domain-specific fine-tuning to train our region-based recognition model. In the next section, we will turn to the testing process of the learned model.

D. Multiscale Testing

In order to reduce the appearance variations in training, the regions extract from pornographic images are normalized to similar scales according to our annotations. However, in the testing phase, we may encounter pornographic images of various scales. To address this issue, we adopt a simple multiscale detection window approach. Specifically, we extract regions at 3 scales for a test image. The smaller two scales are set to 1/2 and 2/3 of the width and height of the whole image, while the large scale is set to the whole image. For the smaller two scales, we extract five regions from four corners and the center, respectively. Therefore, we have 11 regions in total. We resize these regions into 224×224, and if one region is recognized as pornographic, this image is considered as pornographic, while otherwise nonpornographic. In practice, we usually do not need to classify all these regions for a positive image, since once a region is recognized as pornographic, the test process is stopped.

IV. Experiments

In this section, we present a comparative performance evaluation and discussion of the proposed approach. For this, we have collected a large scale dataset from Internet, which covers almost all kinds of pornographic images on Internet, with large variation of backgrounds, scenarios, lightings, and poses. Our most significant results are on this dataset, including the comparisons to traditional methods, commercial systems, and deep learning-based baselines. We also report the running time performance of our method. Furthermore, we validate the effectiveness of the proposed components of the system, and discuss the advantages and limitations of our method.

Researchers have recently reported results on the NPDI pornography dataset [4]. Since the NPDI dataset does not contain the annotations required to train our model, we opt to test on NPDI the model learned from our dataset to demonstrate its generalization capability cross dataset.

A. Datasets

1) Our Dataset: We have collected a large scale dataset consisting of 138 000 pornographic images and 205 000 normal images from Internet. In particular, we download pornographic images from four pornographic websites using a Web crawler software. The downloaded pornographic images are mixed with a small amount of normal images, and we exclude them manually. Our pornographic image set covers almost all kinds of pornographic images on Internet. Overall, we categorize them into three groups, namely, *regular nudity*, *sexual behavior*, and *unprofessional porn*. The *regular nudity* refers to the images of nude people produced by professional photographer. The *sexual behavior* means images depicting sexual behavior, which are also produced by professional photographer. The *unprofessional porn* means pornographic images taken by unprofessional photographer including both nudity and sexual behavior. The unprofessional porn images usually contain complex backgrounds and the image quality is poor.

The normal images of our dataset are also downloaded from Internet, which can be further categorized into three groups,

TABLE I
COMPARISON OF DETECTION RATE (%)

Methods	Porn	Normal	All
Region-of-Interest Method [31]	68.30	61.41	64.85
Improved RoI Method [39]	76.53	66.78	71.65
Bag-of-Feature Method [24]	79.79	71.87	75.83
Improved BoF Method [43]	73.76	82.16	77.96
Commercial System 1 [2]	53.04	**99.16**	76.11
Commercial System 2 [1]	84.37	81.83	83.10
Deep Image-based CNN [28]	88.18	93.66	90.92
Deep Fused CNN [27]	92.67	89.22	90.94
Deep Region-based CNN	95.65	95.04	95.35
Deep Part Detector	92.27	52.49	72.36
Deep MIL	94.15	95.84	94.97
Ours	**98.18**	98.51	**98.35**

namely, *scantily clad people*, *normal people*, and *no people*. In particular, the *scantily clad people* images include bikini, seductive images, and man or baby with bare upper body, which are similar with pornographic images in appearance, and images of *normal people* mean that people in these images are normally dressed, while the images of *no people* consist of various normal images of natural scenery, animals, living goods, and so on, containing no people.

From the pornographic images, we randomly select 33 000 images to annotate their key pornographic contents with bounding boxes. To collect, clean, and annotate these images, we employ six person and spend about two months. In our experiments, we use the annotated 33 000 pornographic images and 100 000 randomly selected normal images as the training set, randomly select 5000 pornographic images and 5000 normal images from the remaining images as the validation set, while the remaining 100 000 pornographic images and 100 000 normal images are used as the test set. Furthermore, to present detailed comparison on different types of images, we roughly split the test images into six groups, i.e., pornographic images: 19 207 *regular nudity*, 37 018 sexual behavior, and 43 775 *unprofessional porn*; and normal images: 68 873 *scantily clad people*, 20 230 normally *dressed people*, and 10 897 *no people* images.

2) NPDI Dataset: The NPDI pornography dataset [4] contains nearly 80 h of 400 pornographic and 400 nonpornographic videos. The videos are already segmented into shots, and on the average, there are almost 20 shots per video. In total, 16 727 key frames selected from the videos, 10 340 normal images, and 6387 pornographic images. However, we find that 1198 of the 6387 pornographic images are incorrectly labeled; hence we remove them from our experiment. Since the NPDI dataset does not have annotations of the key pornographic contents, we do not use it for model training, but for *cross-dataset testing*.

B. Experiment on Our Dataset

On our newly collected large scale dataset, we first compare our method with four traditional methods using shallow low-level features and two commercial systems using deep learning techniques, then we compare it with another five deep learning baselines implemented by ourselves. We use the detection rate and the ROC curve to measure performance.
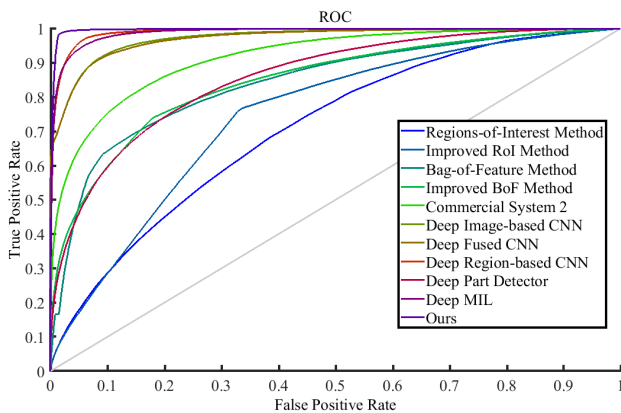
Fig. 1. ROC curve of different methods for pornographic image recognition. Our method significantly outperforms traditional methods and shows some advantage over the in-house baselines using deep representation. In particular, we achieve accuracy of 97.52% TPR at 1% FPR.

The detection rate is computed by the fraction of identified images out of the pornographic/normal images set/subset. Besides comparison to other methods, we also report our running time performance, and validate the proposed system components, and present some discussions about our method.

1) Comparison to Traditional Methods: Traditional works on pornographic image recognition mainly focus on two methodologies: 1) ROI-based methods [31], [39], [45], [47] and 2) BoF-based methods [9], [24], [36], [43]. However, to the best of our knowledge, there is no off-the-shelf recognition systems available for comparison. Hence, we implement two representative ROI-based methods and two BoF methods, respectively. They are the retrieval-based method that extracts color, shape, and skin features on ROI obtained by skin detection [31] and its improved version with a novel skin detection technique [39], and the BoF approach based on Hue-SIFT descriptor [24] and its improved version based on visual attention model [43]. Both of these methods have reported good performances on their small scale datasets.

Rows 1–4 and last row in Table I show the detection rate of the traditional methods and ours on pornographic images, normal images, and all test set. Fig. 1 shows the ROC curve. We can observe that the proposed method significantly outperform the ROI-based method [31] and its improved version [39], as well as the BoF method [24] and its improved version [43]. The results on different types of pornographic images and normal images are shown in Fig. 3. In particular, on the *scantily clad people* subset of normal images, our detection rate is almost twice of the ROI-based method [31].

2) Comparison to Commercial Systems: As pornographic image recognition is very important and useful in practice, some commercial systems have been developed and shared in recently years. We choose two representative commercial systems, which use deep learning techniques and are launched in webpages with URLs [1], [2]. We refer to them as *commercial system 1* and *commercial system 2*, respectively. We note that since these two systems are trained with their own data, a completely objective comparison between them and our system is not possible.

The performance of these two commercial systems on our test set is shown in row 5 and 6 of Table I. The overall performance of these two commercial systems is inferior to that of our method. But we can see that commercial system 1 achieves the best accuracy (99.16%) on our normal image set, comparing to all other methods. In contrast to its high performance on normal image set, it has rather poor prediction accuracy (53.04%) on our pornographic image test set. The commercial system 2 achieves 83.10% accuracy on our test set, which is higher than two traditional algorithms but still have a clear gap with our method.

3) Comparison to Deep CNN Baselines: To the best of our knowledge, there are two deep CNN-based methods for pornographic image recognition [27], [28], which however, have not released their code and training data. To conduct a comprehensive comparison with deep CNN-based baselines, we implement the algorithms described in [27] and [28] and other three CNN-based variations with our training data.

First, since the main idea of [28] is to directly fine-tune the off-the-shelf CNN architecture with their own data, we implement a similar deep CNN version by fine-tuning the GoogLeNet [33] with our training data, and refer it as deep image-based CNN. Second, we reimplement the algorithm in [27] which fuses the recognition results by AlexNet [22] and GoogLeNet [33], and refer it as deep fused CNN. Third, we design a naive version of region-based CNN (similar to [15]) using our bag generation strategy. We assume that if a region's degree of pornography is larger than 0.5, it is labeled as positive and otherwise negative. In this setting, we can train a region-based recognition model using standard supervised learning. We refer to this method as deep region-based CNN. Fourth, based on our annotations, we also train three types of body part detectors with GoogLeNet, including female breast, female sexual organ, and male sexual organ detectors. We resize the annotations for these body parts into 224×224 patches, and fine-tine GoogLeNet with these patches.[3] The trained body part detectors are then used for testing by scanning the test image in a cascaded manner and considering the image as pornographic when one detector has positive response. We refer to this method as *deep part detector*. Finally, we use the standard MIL paradigm to train a region-based detector, without using the weighting strategy. We refer to this method as *deep MIL*.

Rows 7–11 in Table I show the detection rate of these deep CNN-based methods, and Fig. 1 shows the ROC curve. Our method consistently outperforms all deep CNN-based baselines, although most of them can outperform the traditional methods by a large margin. In particular, the deep image-based CNN method outperforms two commercial systems that also employ deep learning techniques. We conjecture that this may be caused by the difference of training data. Furthermore, the deep region-based CNN achieves a significant improvement over the deep image-based CNN method, which confirms the effectiveness of region-based modeling strategy. In addition,

---

[3]The accuracies are 95.02%, 92.72%, and 94.27% for female breast, female sexual organ, and male sexual organ detectors on validation set of their positive and negative patches, respectively.

TABLE II
CROSS-VALIDATION PERFORMANCE (%)

Round	Porn	Normal	All
1	96.68	94.84	95.29
2	97.44	98.34	98.12
3	97.66	97.34	97.42
4	98.18	95.84	96.42
5	96.64	96.32	96.40
6	94.92	95.84	95.61
7	98.04	99.96	99.49
8	97.56	95.36	95.90
9	96.04	99.98	99.26
10	97.58	99.78	99.24
Mean	97.07±0.96	97.40±1.96	97.32±1.54

TABLE III
MULTIPLE TEST PERFORMANCE WITH REDUCED TRAINING DATA (%)

| Round	Porn	Normal	All
1	96.67	95.76	96.22
2	96.86	96.41	96.63
3	97.42	95.28	96.35
4	93.64	93.10	93.37
5	96.80	94.12	95.46
6	93.64	93.10	93.37
7	97.94	97.78	97.86
8	97.22	93.40	95.31
9	96.24	98.18	97.21
10	97384	97.90	97.64
Mean	96.38±1.44	95.50±1.92	95.94±1.51

Fig. 2. Contribution of the components to the performance, including the scale normalization for training regions, the effect of useful context, and the weighting strategy for MIL.

we observe that the body part detector method performs very poorly on the scantily clad images as shown in Fig. 3. This result is consistent with our intuition since these images only have small cover on private parts especially on female breast, which may lead to false positive detection.

4) Cross-Validation Performance: In this section, we verify the generalization ability of our method via cross-validation. In particular, we perform tenfold cross-validation on the training data. Table II gives the validation performance of each round as well as the average accuracy. Our algorithm obtains consistent results among different rounds, and achieves good average performance. With 90% training data, the average cross-validation accuracy is 97.32%, which is very close to the test accuracy with all training data.

Besides cross-validation, we perform two extra experiments to further verify the stability of our algorithm, by changing the setting of training while keeping the test set unchanged. In particular, we first train our model ten times with stochastic gradient descent, and then test these models on the test set, respectively. We surprisingly find that all these models almost produce the same result as reported in Table I. Furthermore, we also test the ten cross-validation models on the full test set, which are trained with 90% training data. The result is shown in Table III. We see that these models show consistency in testing, and the accuracy only drops a little when using the reduced training data.

5) Testing on Gray Images: Almost all the training and testing images in our experiments are color images. But in practice, we may also encounter gray pornographic images. Therefore, we also test our model trained with color images on the black and white copy of our test set. The detection rate on positive images, negative images, and all images are 82.23%, 99.09%, and 90.66%, respectively. The test result is promising, since we did not include any gray images during
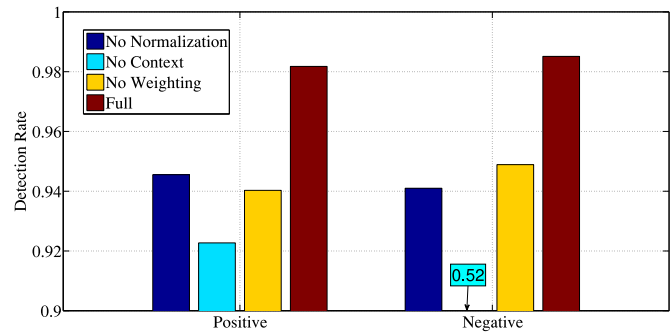
training. In particular, the test result on gray negative images is even better than the counterpart on original color negative images.

6) Running Time Performance: We implement our system with the open-source package Caffe [19], and measure it on an Intel Xeon E5-2630 CPU (2.60 GHz, 12 core) and a GeForce GTX Titan X GPU, respectively. We achieve 5 FPS with CPU, and 55 FPS with GPU. That is, the proposed method is highly efficient and real-time capable when implemented with GPU.

7) Validation of System Components: In this section, we validate the effectiveness of three proposed components of our system, including the normalization for training regions, the inclusion of context information, and the weighting strategy for MIL. To investigate the contribution of the individual component, we conduct a series of experiments on our dataset by removing each of the components in turn while leaving the remaining components in place. Fig. 2 gives the results. In particular, the pipeline that removes useful context can be seen as generic private body part detector, which achieves very poor performance on the negative images. This suggests that the human body parts usually lack distinct patterns without the cue from useful context. We also observe that scale normalization according to our annotations can boost the performance of our method, since it ensures that the positive regions are dominated by pornographic contents to some extent. Lastly, our method clearly benefits from the weighting strategy of MIL, which confirms the importance of incorporating the region's degree of pornography into MIL.

8) Discussions: In this section, we present discussions about the advantages and limitations of our method, by analyzing our performance on different types of images.

We summarize the performance of our method and other comparative methods on different types of test images in Fig. 3. Our method performs extremely consistently on different images, while other methods rise and fall greatly. In particular, our method shows obvious advantages over all other methods on the *unprofessional porn* images. This part of pornographic images mainly come from unprofessional photographs in daily life, which exhibit large variations in background, scale, lighting, and people pose. This suggests that our method can capture the essential pornographic contents of these images to some extent, while other methods

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.
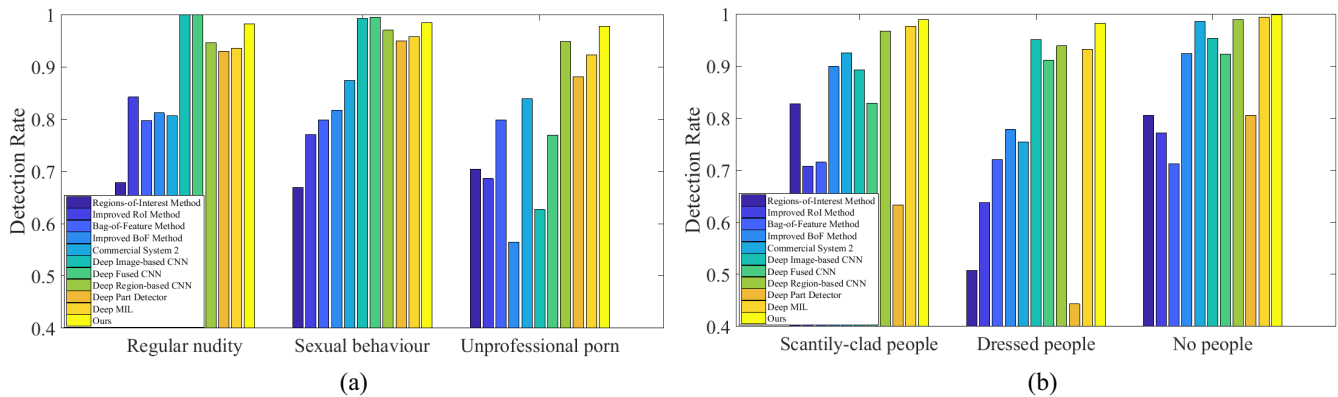
8

IEEE TRANSACTIONS ON CYBERNETICS

Fig. 3.   Illustration of results (detection rate) on different types of (a) pornographic images and (b) normal images.

(e.g., deep image-based CNN) may be effected by the large background of these images. We notice that our performance on scantily clad images are even better than that on normally dressed images. This seems a little strange, and we conjecture that it is because the deep representations based on useful context information have the capability to discriminate exposed private body parts and private body parts with a small cover, while the normally dressed images in life have larger variations in clothe styles, poses, and scenes compared to scantily clad images, which may bring in challenges for recognition.

Many of our true positive images exhibit large variations of lighting, pose, scale, and occlusion. On the other hand, we note that false detections are likely to happen in the local regions that look like exposed female breast, while actually they are exposed male breasts. In addition, since we consider the exposure of private body parts or sex acts as the criteria for pornographic images, our current system cannot properly distinguish some art works with exposed private part from pornographic images. But fortunately, such nude art works are rather a small part of Internet images.

Furthermore, we also find many challenging true negative images on our dataset, including seductive images, bikini girls, men with naked upper body, and naked babies. These images, which typically show similar appearance with pornographic images, can to some extent validate the capability of our method to distinguish pornography from some related concept, such as seduction and nudity. In addition, the typical false negative images of our method often show very little private body parts such that they can hardly be correctly detected.

C. Cross-Database Experiment on NPDI Dataset

On the NPDI Pornography dataset, we test our recognition model trained on our dataset to demonstrate its generalization capability across datasets. The classification performance is evaluated using the standard metric for this dataset, the mean average precision (MAP). Our model achieves good performance, reaching a MAP of 0.975, while the best performance in literature on this database is 0.964 (MAP) [3].

Table IV shows detailed comparison between our method and many other methods which are also trained on our dataset. Our method consistently outperforms four traditional methods and five deep learning baselines. We note that our model

TABLE IV
COMPARISON ON NPDI BY MAP

Methods	MAP
Region-of-Interest Method [31]	0.712
Improved RoI Method [39]	0.756
Bag-of-Feature Method [24]	0.811
Improved BoF Method [43]	0.823
Deep Image-based CNN [28]	0.874
Deep Fused CNN [27]	0.919
Deep Region-based CNN	0.878
Deep Part Detector	0.870
Deep MIL	0.860
BossaNova [3]	0.964
Ours	**0.975**

performs well, although there are some differences between the styles of the training pornographic images from our database and the test pornographic images from NPDI dataset.

V. CONCLUSION

In this paper, we follow an MIL approach to address the problem of pornographic image recognition. In particular, we exploit the annotations of the key pornographic contents in the training phase. These annotations help us to generate a variety of regions with different degree of pornography, and to define a quantitative measure of these regions' degree of pornography. Based on these, we formulate the recognition task as a weighted MIL problem under the CNN framework, resulting a robust region-based recognition model. We collect a large scale dataset consisting of 138K pornographic images and 205K normal images from Internet, and our system demonstrates strong performance on this dataset, significantly outperforms traditional methods and several deep learning baselines. Additionally, the proposed method is highly efficient and real-time capable when implemented with GPU.

REFERENCES

[1] (2016). *NSFW Link*. [Online]. Available: https://github.com/yahoo/open_nsfw

[2] (2016). *Qcloud Link*. [Online]. Available: http://www.qcloud.com/event/pd

[3] S. Avila, N. Thome, M. Cord, E. Valle, and A. de Albuquerque Araújo, "Pooling in image representation: The visual codeword point of view," *Comput. Vis. Image Understand.*, vol. 117, no. 5, pp. 453–465, 2013.

[4] S. Avila, N. Thome, M. Cord, E. Valle, and A. de Albuquerque Araújo, "BOSSA: Extended bow formalism for image classification," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, 2011, pp. 2909–2912.

[5] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[6] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, pp. 913–939, Aug. 2004.

[7] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 2409–2416.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2005, pp. 886–893.

[9] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *Proc. 19th Int. Conf. Pattern Recognit. (ICPR)*, 2008, pp. 1–4.

[10] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, 1997.

[11] M. M. Fleck, D. A. Forsyth, and C. Bregler, "Finding naked people," in *Computer Vision—ECCV*. Heidelberg, Germany: Springer, 1996, pp. 593–602.

[12] D. A. Forsyth and M. M. Fleck, "Identifying nude pictures," in *Proc. 3rd IEEE Workshop Appl. Comput. Vis. (WACV)*, 1996, pp. 103–108.

[13] D. A. Forsyth and M. M. Fleck, "Automatic detection of human nudes," *Int. J. Comput. Vis.*, vol. 32, no. 1, pp. 63–77, 1999.

[14] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 264–277, Sep. 1999.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 580–587.

[16] R. Hong *et al.*, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.

[17] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank, "Recognition of pornographic Web pages by classifying texts and images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1019–1034, Jun. 2007.

[18] S.-W. Jang, Y.-J. Park, G.-Y. Kim, H.-I. Choi, and M.-C. Hong, "An adult image identification system based on robust skin segmentation," *J. Imag. Sci. Technol.*, vol. 55, no. 2, pp. 1–10, 2011.

[19] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[20] F. Jiao, W. Gao, L. Duan, and G. Cui, "Detecting adult image using multiple features," in *Proc. Int. Conf. Info Tech Info Net (ICII)*, vol. 3. Beijing, China, 2001, pp. 378–383.

[21] S. Kim, H. Min, J. Jeon, Y. M. Ro, and S. Han, "Malicious content filtering based on semantic features," in *Proc. 2nd Int. Conf. Interact. Sci. Inf. Technol. Culture Human*, 2009, pp. 802–806.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[23] D. Li, N. Li, J. Wang, and T. Zhu, "Pornographic images recognition based on spatial pyramid partition and multi-instance ensemble learning," *Knowl. Based Syst.*, vol. 84, pp. 214–223, Aug. 2015.

[24] A. P. B. Lopes, S. E. de Avila, A. N. A. Peixoto, R. S. Oliveira, and A. de Albuquerque Araüjo, "A bag-of-features approach based on hue-SIFT descriptor for nude detection," in *Proc. 17th Eur. Signal Process. Conf.*, 2009, pp. 1552–1556.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[26] L. Lv *et al.*, "Pornographic images detection using high-level semantic features," in *Proc. 7th Int. Conf. Nat. Comput. (ICNC)*, vol. 2, 2011, pp. 1015–1018.

[27] M. Moustafa, "Applying deep learning to classify pornographic images and videos," *arXiv preprint arXiv:1511.08899*, 2015.

[28] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," *Neurocomputing*, vol. 210, pp. 283–293, Oct. 2016.

[29] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[30] C. X. Ries and R. Lienhart, "A survey on visual adult image recognition," *Multimedia Tools Appl.*, vol. 69, no. 3, pp. 661–688, 2014.

[31] J.-L. Shih, C.-H. Lee, and C.-S. Yang, "An adult image identification system employing image retrieval technique," *Pattern Recognit. Lett.*, vol. 28, no. 16, pp. 2367–2374, 2007.

[32] L. Sui, J. Zhang, L. Zhuo, and Y. C. Yang, "Research on pornographic images recognition method based on visual words in a compressed domain," *IET Image Process.*, vol. 6, no. 1, pp. 87–93, Feb. 2012.

[33] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[34] J. Tang, H. Li, G.-J. Qi, and T.-S. Chua, "Image annotation by graph-based inference with integrated multiple/single instance representations," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 131–141, Feb. 2010.

[35] J. R. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[36] A. Ulges and A. Stahl, "Automatic detection of child pornography using color visual words," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2011, pp. 1–6.

[37] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.

[38] Y. Xiao, B. Liu, Z. Hao, and L. Cao, "A similarity-based classification framework for multiple-instance learning," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 500–515, Apr. 2014.

[39] J.-J. Yu and S.-W. Han, "Skin detection for adult image identification," in *Proc. Int. Conf. Adv. Commun. Technol.*, 2014, pp. 645–648.

[40] A. Zaidan, H. A. Karim, N. Ahmad, B. Zaidan, and A. Sali, "A four-phases methodology to propose anti-pornography system based on neural and Bayesian methods of artificial intelligence," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 28, no. 1, 2014, Art. no. 1459002.

[41] P. A. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1417–1424.

[42] C. Zhang and P. A. Viola, "Multiple-instance pruning for learning efficient cascade detectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1681–1688.

[43] J. Zhang, L. Sui, L. Zhuo, Z. Li, and Y. Yang, "An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain," *Neurocomputing*, vol. 110, no. 8, pp. 145–152, 2013.

[44] K. Zhang and H. Song, "Real-time visual tracking via online weighted multiple instance learning," *Pattern Recognit.*, vol. 46, no. 1, pp. 397–411, 2013.

[45] Q.-F. Zheng, W. Zeng, W.-Q. Wang, and W. Gao, "Shape-based adult image detection," *Int. J. Image Graph.*, vol. 6, no. 1, pp. 115–124, 2006.

[46] Z.-H. Zhou, "Multi-instance learning: A survey," Dept. Comput. Sci. Technol., Nanjing Univ., Nanjing, China, Rep., 2004.

[47] Q. Zhu, C.-T. Wu, K.-T. Cheng, and Y.-L. Wu, "An adaptive skin model and its application to objectionable image filtering," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 56–63.

[48] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.
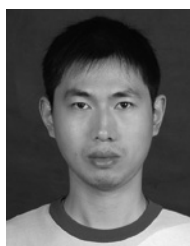
Xin Jin received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, 2012, and 2017, respectively.

He is currently a Computer Vision Researcher with Megvii Research Nanjing (Face++), Nanjing. His current research interests include computer vision and deep learning, especially focusing on face landmark detection and general object detection.

Yuhui Wang, photograph and biography not available at the time of publication.

Xiaoyang Tan received the B.Sc. and M.Sc. degrees in computer applications from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1993 and 1996, respectively, and the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, in 2005.

In 1996, he was an Assistant Lecturer with NUAA. From 2006 to 2007, he was a Post-Doctoral Researcher with the Learning and Recognition in Vision team, INRIA Rhone-Alpes, Grenoble, France. His current research interests include face recognition, machine learning, pattern recognition, and computer vision. He has authored or co-authored over 20 scientific papers in the above areas.