大量的裸露肤色信息这一事实, 通过肤色检测来判断是否不良图片. RGB 颜色空间、HSV 和 HIS 颜色空间以及亮度和色度分开存储的 YCbCr 颜色空间[2]都曾被利用以检测肤色. 也有研究人员使用多重颜色空间定义肤色模型, 并加入纹理检测以更准确地提取出肤色区域, 然后利用裸露肤色区域的像素比例[3]、肤色连通域的数目、位置、形状[4]等特征作为 SVM 等分类器的输入, 输出图片是否不良图片. 这类基于肤色特征和人体检测的方法存在的主要问题是, 人工选择的肤色特征模型总是带有一定的偏差, 在现实中不同种族的肤色各不相同, 相同的肤色在不同的光照下也会呈现不同的结果, 再加上自然界中存在与肤色类似的物体, 这类基于低层次语义特征进行相似度匹配的方法误检率较高. 其它方法包括结合不同的人体识别技术来提高人体检测的准确率, 如人体动态姿势识别的多层联合模型[5], 以及在具有复杂背景的图像中结合图像区域检测技术[6]进行更准确的场景分割与识别.

基于深度学习的方法把特征提取与分类器放在一个模型里统一起来, 特征提取是通过大量的训练数据自动学习得到. 2012 年文献[7]提出 AlexNext 网络结构模型, 2015 年文献[8]提出的 ResNet 模型, 都大幅提升了图片的分类准确率. 因此文献[9]提出 AGNet 模型, 利用卷积神经网络 (CNN) 解决不良图片的识别问题, 取得了较好的效果. 文献[10]提出先通过人脸识别和肤色检测的 Coarse Detection 方法过滤一部分图片, 再通过 CaffeNet 进行分类. 文献[11]提出将多个 CNN 的输出进行加权综合的方法. 由于 CNN 的参数多, 必须依靠大规模的训练数据, 这在色情图片识别领域不容易获得. 这些方法通过 CNN 进行二分类, 没有考虑到色情图片类别的多样性, 在面对实际应用中具有多样性特征的色情图片时, 网络模型的性能和识别准确率往往会下降.

本文利用深度残差网络算法, 结合不良图片的多样性特征, 提出一种基于多分类和 ResNet 的不良图片识别框架. 通过更细粒度的分类, 将二分类的色情图片识别问题转化为多分类问题, 再依据输出分值判断为良性或不良图片. 在训练时采用依据测试结果反馈补充边缘案例的训练策略, 让模型挖掘更优质的特征. 在测试时采用一种单边滑动窗口的方法, 以降低不同图片尺度带来的影响. 实验结果表明该方法能够有效提高不良图片识别的性能.

1 基于多分类和 ResNet 的不良图片识别框架

深度学习算法能够自动提取图片的抽象特征和语义特征, 这可以避免手动设计的特征模型与真实特征有差异的弊端. 本文设计的基于多分类的 ResNet 不良图片识别框架如图 1 所示. 首先从网上获取良性图片和不良图片构建数据集, 并对图片标记正确的类别标签, 再通过单边滑动窗口方法将图片分为多个图片碎片, 将这些碎片送入 ResNet 进行分类, 最后对分值进行统计和计算并结合合适的阈值划分为良性图片或不良图片.

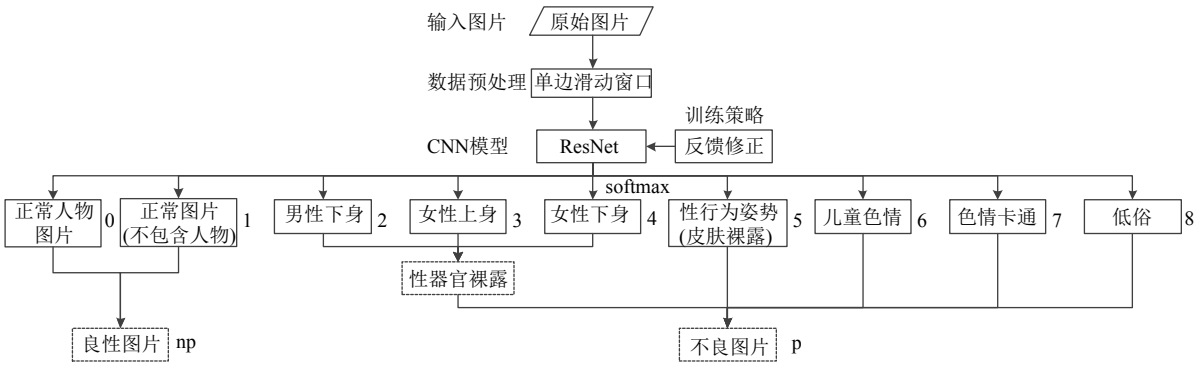

图 1 不良图片识别框架图

1.1 数据预处理

CNN 在处理不同尺度的图片时多采用滑动窗口机制, 基于网络模型的输入大小 (本文中为 224×224) 作为滑动窗口. 在不良图片识别领域, 由于敏感内容只是图片中的一小部分, 将图片直接 resize 到网络的输入大小会导致一定的尺度比例损失. 而采用普通的滑

动窗口机制, 在两条边上以较小步长 (一般为 2) 滑动会产生较多的图片碎片, 导致时间性能的下降和造成一定的误判. 本文通过一种单边滑动窗口的机制进行数据的预处理, 先将待处理图片中最短边 resize 为 224, 另一边进行同比例的 resize, 然后用 224×224 的窗口在图片上沿最长边滑动, 步长设置为 50. 实验表明, 大多数图片的长宽比在 1:1 到 2:1 之间, 很少有图片超过 2:1 的长宽比, 采用 50 的步长可以将滑动窗口产生的图片碎片保持在 3 到 5 个之间, 在保留了图片的比例信息的同时有效减少数据量. 在将每一个图片碎片输入模型处理的过程中, 如果按照阈值被分类为不良图片, 则不再继续滑动, 算法流程如图 2 所示.

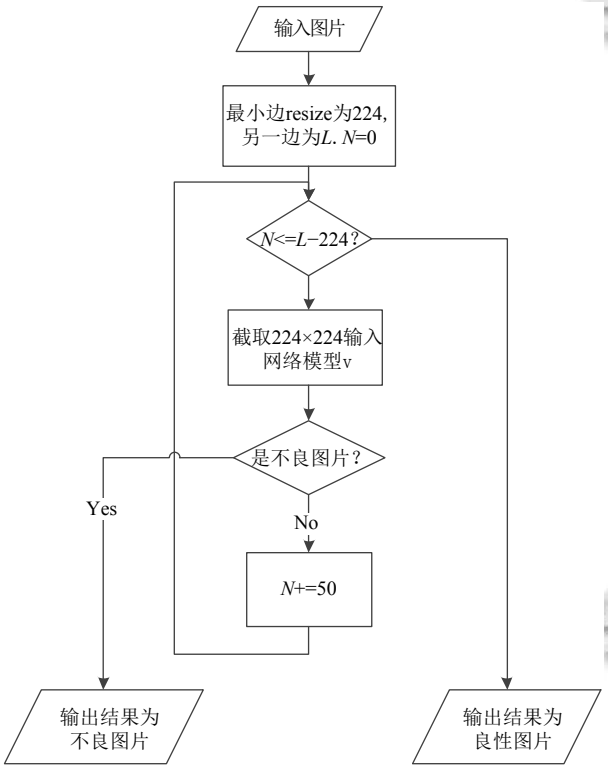

图 2 单边滑动窗口方法

1.2 基于 ResNet 的细粒度分类方法

1.2.1 深度残差网络

AlexNet[7]、VGGNet[12]和 GoogleNet[13]已被证明在图像分类任务上可以获得良好的效果. 本文采用的是 ResNet[8], ResNet 是何凯明于 2015 年提出的 CNN 结构模型, 该方法以 152 层的网络模型在 ILSVRC2015 上获得第一名, 将错误率降低到了 3.75%. ResNet 的主要优点是可以利用更深层次的网络解决训练误差随网络层数的增加而增大的问题. 为了解决该问题, ResNet 对传统的平原网络结构进行了调整, 其关键结构是将基本的网络单元增加了一个恒等的快捷连接 (如图 3 所示). 图中 $H(x)$ 为理想映射, $F(x)$ 为残差映射, $H(x)=F(x)+x$. 通过将拟合目标函数 $H(x)$ 转变为拟合残差函数 $F(x)$, 把输出变为拟合和输入的叠加, 使得网络对输出 $H(x)$ 与输入 x 之间的微小波动更加敏感.

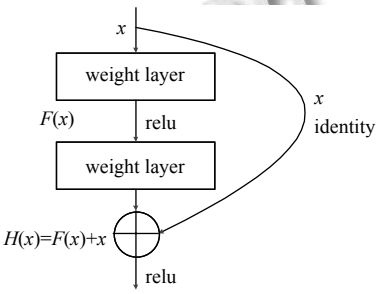

图 3 添加了快捷连接的残差网络单元

本文构建的是 50 层 ResNet 结构模型, 并将每一层的参数减半, 以减少训练时间和在分类准确性与时间性能上做出平衡. 本文利用深度学习框架 Caffe 构建网络模型, 构建的模型共有约 590 万个参数, 网络中的神经元连接总数约为 10.68 亿. 相比于 AlexNet 的约 6000 万个参数, 该模型在网络层数、需训练的参数个数上优势明显. 由于框架分类输出为 9 类, 故将最后全连接层的神经元个数设置为 9, 并将权值学习率和偏置学习率调整为 10 倍. 模型的输入为 224×224 大小的图片, 第一个卷积层 conv_1 的参数是 64 个 7×7 的卷积核, 卷积核的步长为 2. 每一个卷积层之后都设置 BatchNorm 层, 以增加模型的容纳能力. 激活函数使用 Relu, 并通过最大池化层 maxpool 进行下采样. 最后一层是 softmax 回归层, 用于输出图片被分为某一类的概率. softmax 回归是 logistic 回归的一般形式, 其数学公式为公式 (1) 所示, 其中 k 为类别数, 当 k=2 时 softmax 退化为 logistic. 本文实验中训练的模型有 9 个类别, 故 k=9.

$$h_\theta(x^{(i)})=\begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix}=\frac{1}{\sum_{j=1}^{k} e^{\theta_j^{\mathrm{T}}x^{(i)}}}\begin{bmatrix} e^{\theta_1^{\mathrm{T}}x^{(i)}} \\ e^{\theta_2^{\mathrm{T}}x^{(i)}} \\ \vdots \\ e^{\theta_k^{\mathrm{T}}x^{(i)}} \end{bmatrix} \quad (1)$$

1.2.2 细粒度分类方法

具有不同特征的图片都可以称为不良图片, 但简单的将其划分为同一类别会让网络模型在学习其高层语义特征时产生困惑, 导致在处理某些图片时分类混乱, 降低网络性能. 因此基于对不良图片多样性特征的分析, 将 ResNet 的输出分为更细粒度上的 9 个分类, 以便更准确地提取出不良图片的高层语义特征. 这些分类的定义如下:

类别 0: 良性图片中的正常人物类图片;

类别 1: 良性图片中的正常类图片 (无人物);

类别 2: 男性下身类图片, 属于不良图片中特征明显的性器官裸露类;

类别 3: 女性上身类图片, 属于不良图片中特征明显的性器官裸露类;

类别 4: 女性下身类图片, 属于不良图片中特征明显的性器官裸露类;

类别 5: 不良图片中的性行为姿势类图片, 一般包含大量的皮肤裸露;

类别 6: 包含儿童色情类的不良图片;

类别 7: 包含色情信息的卡通类图片;

类别 8: 不良图片中的低俗类, 包含裙底偷拍、人体内衣敏感部位特写等.

其中良性图片和不良图片定义为抽象类别, 不良图片中性器官裸露类也定义为抽象类别. 这种分类方法基本涵盖了不良图片的内容范围, 并使得同一类别内的特征尽量统一, 增大类别间的特征差异性, 以提升网络模型的分类准确率. 实验发现, softmax 分类器在类别之间的特征互斥时效果最好, 因此使用类别特征明显的训练图片可以增加模型的准确率.

在网络模型输出各类别的分值后, 分类为良性图片和不良图片的方法如下:

1) 取分值最大的类别 n 和分值 s;

2) 如果 n=0 或 1, 分类为良性图片;

3) 如果 n 是 2～4, s=s×1.2;

4) 如果 n 是 5～8, s=s×0.92;

5) 如果 s>=0.85, 分类为不良图片.

其中 2、3、4 这三类是主要的不良图片构成部分, 也是特征最明显的类别. 经过试验分析, 在统计分值时对这三类硬色情图片的分值乘以系数 1.2, 可以增强过滤的准确率. 低俗类的属性特征相对模糊, 其边界难以与性感图片区分开. 性行为姿势类、儿童色情类的色情标准同样具有一定的模糊性. 卡通漫画本身的描述方法就具有夸张性, 卡通色情类的判定也应比正常色情类弱. 将 5、6、7、8 类软色情图片的分值乘以 0.92, 进行一定的弱化, 可以减少定义模糊的图片类型的干扰. 将最后得到的分值与阈值 0.85 进行比较, 当大于等于 0.85 时分类为不良图片.

1.3 基于反馈的训练策略

所有的良性图片和不良图片都是基于上述分类特征从网上下载得到, 并被分为训练数据集和测试数据集. 由于实际中无法构建百万级的大规模不良图片数据集, 因此首先利用 ImageNet 1000 数据集进行预训练, 以学习到可以在接下来的训练中利用的参数权值, 减少进一步训练所需的时间. 然后再在训练数据集上进行训练, 并在测试数据集上测试模型的分类效果, 直至网络收敛. 采用了一种反馈修正的训练策略, 在每进行 20 个 epoch 之后, 从每个类别中随机挑选 100 张图片, 分别测试模型的分类准确率. 对准确率小于 0.9 的类别, 定向增加其训练样本的容量, 包括增加与测试样本具有相似特征的图片以及不同肤色的图片等边缘案例, 再继续进行训练. 多次重新设计和构建训练数据集直到模型可以挖掘出更优质的特征. 训练过程如图 4 所示.

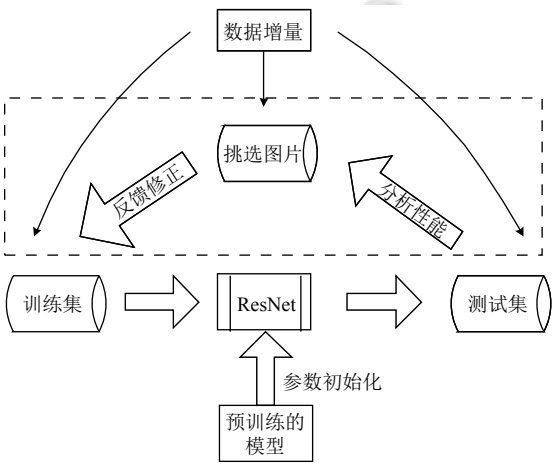

图 4 基于反馈修正的训练策略

2 实验

2.1 已有数据集

不良图片识别具有一定的法律特殊性, 大部分已有的研究都是基于研究人员自己构建的数据库. 而对这些数据库的描述往往不够清晰, 只包含基本的色情

图片和非色情图片的数目 (如表 1), 图片的来源往往都描述为从网上下载得到, 对数据集中图片的多样性特征、难易程度、类型等可能影响识别准确率的因素往往没有提到. 目前缺少一个公开的具有一定标准的不良图片数据库, 因而难以对不同的不良图片识别方法进行比较和评估.

表 1 不同研究方法的数据集描述

色情图片识别方法	数据集描述	图片来源
AGNet[9]	色情图片 8363 张, 非色情图片 8363 张	Google NPDI
Coarse Detection+CaffeNet[10]	色情图片 8000 张, 非色情图片 11 000 张	网上下载得到
有师监督多示例学习 (SD-MIL)[14]	色情图片 155 000 张, 非色情图片 222 000 张	网上下载得到
CaffeNet[15]	色情图片 13 300 张, 非色情图片 35 300 张	网上下载得到
Mixture CNN[16]	色情图片 41 154 张, 非色情图片 40 152 张	网上下载得到
CNN Ensemble[11]	色情图片 10 997 张, 非色情图片 38 832 张	网上下载得到

2.2 训练集构建

本文基于更细粒度的特征分类方法构建了相应的训练数据集和测试数据集. 在挑选图片时, 主要选择特征突出、背景简单或无背景的图片, 以降低过拟合问题的影响. 不同于一般的图片分类任务, 不良图片识别具有很高的主观性和复杂多样性. 例如, 同样的图片, 不同的人可能会划分为不同的类别 (性感类或色情类), 甚至同一人在不同的时间对相同图片的分类也可能不一致. CNN 的最终分类效果对训练数据的要求较高, 特征不明显或标签错误的训练图片将导致网络不能提取出有效的特征. 为了构建更优质的训练数据集, 本文通过人工筛选出有效的图片并对图片标记正确的标签.

筛选时在每一类别中涵盖容易和困难部分, 以覆盖多样化的场景和增加分类时的泛化能力. 例如, 正常图片 (不包含人物) 类别的容易部分覆盖了不同主题的图片, 如风景、汽车、动物、建筑物等; 困难部分则包含不良图片识别容易出错的点, 如类肤色的木头等. 而对于正常的人物图片, 其容易部分也会覆盖肤色、种族等多样性, 困难部分则覆盖含大片皮肤裸露的运动 (游泳、拳击、摔跤) 以及哺乳、比基尼、内衣、给婴儿或小孩洗澡等图片. 数据集的部分图片样例如图 5 所示.

图 5 数据集的部分样例

本文构建的数据集分布如表 2 所示.

表 2 数据集类别分布

类别	良性图片		不良图片						
	正常人物	正常非人物	男性下身	女性上身	女性下身	性行为姿势 (皮肤裸露)	儿童色情	色情卡通	低俗
样本数目	12 000	6546	3834	3829	3796	2760	2334	3542	2689

2.3 实验测试分析

本文实验主要基于一块显存为 12 GB 的 GeForce GTX Titan X 显卡, 实验平台操作系统为 ubuntu 16.04, 网络模型通过深度学习开源工具 Caffe 进行搭建, 编程语言使用的是 Python 2.7.

从构建的数据集中随机挑选一定的图片分别用作训练集和测试集, 图片数目如表 3 所示. 为了进一步增大数据集的容量, 采用多种数据增量技术. 常用的随机裁切方法可以产生较多的图片碎片, 但并不适用于不良图片, 因为这些碎片可能已经改变了其类别属性, 产生标签错误. 因此采用不改变其分类特征的方法, 首先通过旋转和翻转, 每次将图片旋转 45 度, 再进行水平和垂直翻转, 得到 32 张图片. 其次向图片增加随机噪声, 并在 0.9～1.1 的范围内随机轻微修改图片亮度, 得到 5 张图片. 最后采用的方法是高斯模糊处理.

模型在 ImageNet 1000 数据集上预训练之后, 再在训练集上进行训练. 权值的基础学习率设置为 0.01, 动量为 0.9, 权值衰减为 0.0005, 采用前文描述的反馈修

正策略训练直至网络收敛.

表 3 构建的训练数据集和测试数据集图片数目

	训练集	测试集	总数
不良图片	18 322	4462	22 784
良性图片	14 390	4156	18 546
总数	32 712	8618	41 330

训练好的模型文件大小约为 23 MB, 与 CaffeNet 的 230 MB 相比约为 1/10. 这是因为 ResNet 模型的参数数量相比 CaffeNet 模型少了很多. 训练好的模型的第一个卷积层 conv_1 的 64 个卷积核如图 6 所示. 这些卷积核可以提取到图片的角度、边缘和颜色等特征, 再将这些特征送入高层网络进行进一步的特征提取. 一般而言, 第一个卷积层的卷积核越清晰越能提取出优质的特征.

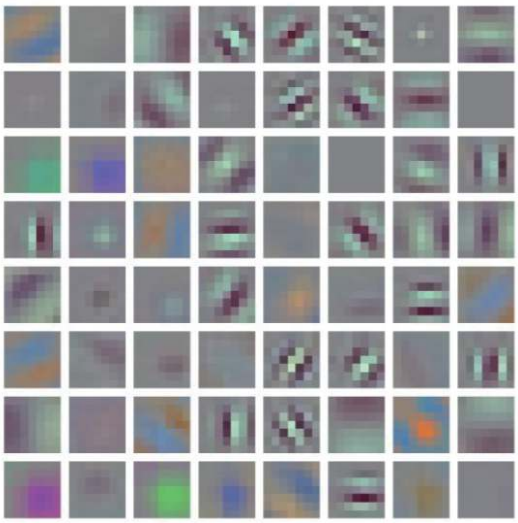

图 6 conv_1 训练好的卷积核参数

2.3.1 识别准确率分析

本文对不良图片识别的分类结果采用准确率 *AUC* 来表示, *T* 为测试集上分类正确的样本数, *P* 为测试集的总样本数目, 则 *AUC* 为:

$$AUC = \frac{T}{P} \tag{2}$$

对图片进行测试, 过滤阈值可以根据具体应用环境来设置. 实验中, 通常情况下最后输出分值 s 大于 0.85 时图片是不良图片的概率较大, 小于 0.3 时是良性图片的概率较大. 在一般场景下可将阈值设定为 0.85. 该方法获得了较高的识别准确率. 与其他方法的准确率对比如表 4 所示.

表 4 准确率 AUC 对比

方法	准确率 (%)
ORB+HSV[17]	91.03
Coarse Detection+CaffeNet[10]	95.20
SD-MIL[14]	96.33
Mixture CNN[16]	95.36
CNN Ensemble[11]	93.84
基于多分类和 ResNet 的框架	97.53±0.5

基于 ORB+HSV 的方法利用了 HSV 颜色空间, 结合构造的具有鲁棒性的特征描述符构建词袋模型, 最后通过 SVM 进行分类. 该方法对于肤色信息有较好的独立性, 但对图片特征的拟合均有一定的误差, 准确率较低. CaffeNet、CNN 和多示例学习 (MIL) 对光照、图片尺度、清晰度和图片类型的变化具有较强的鲁棒性, 可以达到较高的准确率, 但这些方法忽视了色情图片的多样性特征. 本文方法充分考虑了多样性特征的影响, 所以在类别丰富的数据集上表现更好. 准确率波动幅度较大的原因可能是类别 5～8 的正样本数目相对较少, 保持训练正样本的类别比例约为 1:1 可带来更好的效果.

2.3.2 时间效率分析

在对不良图片的识别过滤处理中, 算法的时间性能也很重要. 本文方法对单张图片的处理时间小于 40 ms, 与其它方法的 80 ms 以上相比有很大的提升. 时间性能对比如表 5 所示. 基于手动特征设计的 ORB+HSV 算法花费时间较多, 基于深度学习的方法通过 GPU 进行运算, 时间性能相对较高, 本文方法的网络模型参数较少, 可以达到更高的时间效率.

表 5 时间性能对比

方法	时间 (ms)
ORB+HSV[17]	769
Coarse Detection+CaffeNet[10]	84
SD-MIL[14]	112
Mixture CNN[16]	96
CNN Ensemble[11]	92
基于多分类和 ResNet 的框架	35

3 结论

本文提出一种基于多分类和 ResNet 的不良图片识别框架, 结合更细粒度的分类特征, 将不良图片识别转化为多分类任务. 通过反馈修正的训练策略和单边滑动窗口的测试方法, 根据具体环境设置阈值来划分

良性和不良图片. 实验结果表明, 该方法可以以较少的测试时间获得较高的不良图片识别准确率. 但该方法所需训练时间仍较长, 低俗类等的边界条件仍有一定的模糊性, 接下来将探讨划分更多子类的可能性, 以进一步优化算法的性能.

参考文献

1 凡阿杰. 网络视频色情信息的检测技术研究与应用[硕士学位论文]. 北京: 北方工业大学, 2017.

2 裴向杰, 唐红昇, 陈鹏. 融合 YCbCr 肤色分割的不良图像检测算法研究. 计算机技术与发展, 2015, 25(12): 80–84, 90.

3 王景中, 周靖. 基于比例特征的网络不良图像过滤算法研究. 计算机工程与科学, 2016, 38(3): 514–519. [doi: 10.3969/j.issn.1007-130X.2016.03.018]

4 黄杰, 史啸. 一种基于人体裸露皮肤形状的不良图像过滤系统. 东南大学学报 (自然科学版), 2014, 44(6): 1111–1115.

5 Ding M, Fan GL. Multi-layer joint gait-pose manifold for human motion modeling. IEEE Transactions on Cybernetics, 2015,45(11):2413–2424.[doi:10.1109/TCYB.2014.2373393]

6 Shao H, Yu TS, Xu MJ, *et al*. Image region duplication detection based on circular window expansion and phase correlation. Forensic Science International, 2012, 222(1–3): 71–82. [doi: 10.1016/j.forsciint.2012.05.002]

7 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3098997]

8 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

9 Moustafa M. Applying deep learning to classify pornographic images and videos. arXiv preprint arXiv: 1511.08899, 2015.

10 Zhou KL, Zhuo L, Geng Z, *et al*. Convolutional neural networks based pornographic image classification. 2016 IEEE Second International Conference on Multimedia Big Data. Taipei, China. 2016. 206–209. [doi: 10.1109/BigMM.2016.29]

11 Huang Y, Kong AWK. Using a CNN ensemble for detecting pornographic and upskirt images. 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). Niagara Falls, NY, USA. 2016. 1–7.

12 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Computer Science, 2014: 1–14. [doi: 10.5121/csit.2014.41000]

13 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9. [doi: 10.1109/CVPR.2015.7298594]

14 Wang YH, Jin X, Tan XY. Pornographic image recognition by strongly-supervised deep multiple instance learning. 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA. 2016. 4418–4422. [doi: 10.1109/ICIP.2016.7533195]

15 Nian FD, Li T, Wang Y, *et al*. Pornographic image detection utilizing deep convolutional neural networks. Neurocomputing, 2016, 210: 283–293. [doi: 10.1016/j.neucom.2015.09.135]

16 Connie T, Al-Shabi M, Goh M. Smart content recognition from images using a mixture of convolutional neural networks. IT Convergence and Security 2017. Springer. Singapore. 2018. 11–18.

17 Zhuo L, Geng Z, Zhang J, *et al*. ORB feature based web pornographic image recognition. Neurocomputing, 2016, 173: 511–517. [doi: 10.1016/j.neucom.2015.06.055]