图片内容鉴黄算法综述

概述

在互联网时代，我们能够通过互联网传输，获取海量的信息。这些信息以文字，音频，图像，视频等形式呈现给广大的用户。但是，对于广大用户而言，这些信息并不一定都是有效信息。其中，包含了大量的垃圾数据（如：垃圾邮件，推广广告）。甚至包含了一些违反当地政策法规的信息（如：色情，暴力，涉恐，涉政的文字图片影音等）。如何高效的将上述垃圾数据，违规信息从海量的互联网数据中过滤掉，已经成为一个亟待解决的问题。

在上述众多类型的违规信息中，涉黄信息是往往是最常见的。同时，对于涉黄数据的过滤也是业界关注的比较早。本文只关注于涉黄图片的检测和识别这一小领域。对于其他表现形式的涉黄信息（如：文字，音频，视频等）的识别不作讨论。

前言

互联网数据中的违规信息根据其类型，可以分为色情，暴力，涉恐，涉政等不同类型。每种类型也有不同的具体的表现形式，如：文字，图片，视频，音频等。这类信息习惯上也被称为 NSFW（Not Suitable For Work），即不适合上班时间浏览的网络内容。这类内容的定义其实比较主观，而且在不同国家或地区，甚至于针对不同的互联网用户，它的定义标准差别非常大（例如：在一些严苛的穆斯林国家，图片中的女性不把自己完全遮住，这张图片就是违规的色情图片。而在一个世俗国家，女性泳装图片是正常图片而允许浏览；成人图像对于成年人是合规的允许浏览的，但是对于未成年人又变成不合规的）。如此种种，使得 NSFW 数据的检测在一开始就困难重重。

本文所关注的涉黄图片的识别【图 1】，只是违规信息识别的一个子集。也面临着同样的问题。就目前收集到的信息，还有没有一个统一的标准，判定一张图片是否真的就涉黄。不论在 paper 里，还是在一些已经在商用的图片鉴黄产品里，图片涉黄的判别标准要么是模糊的，要么干脆是不予公开。

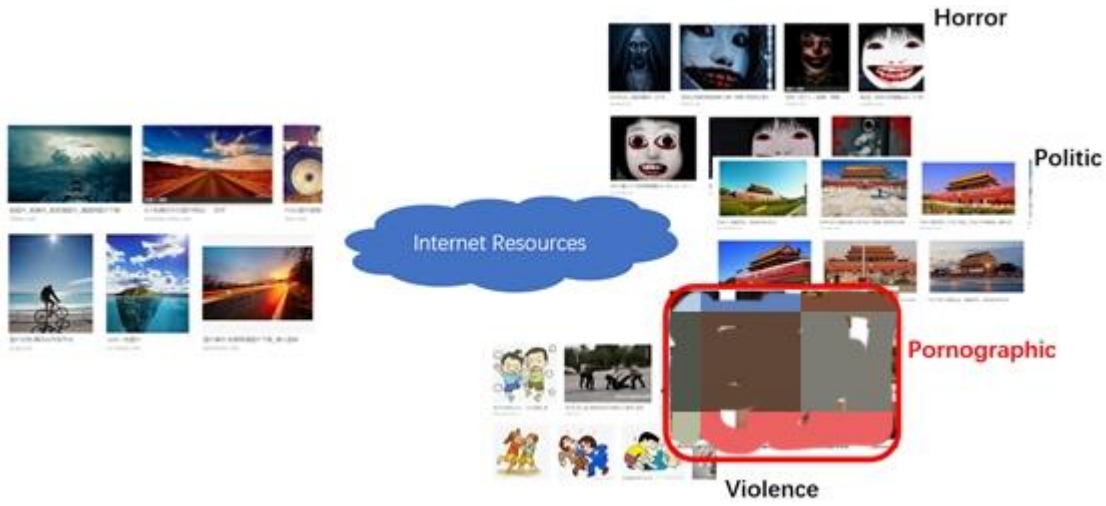


图 1 违规信息识别

不单是涉黄图片评价标准的不一致，不同 paper 中，使用的涉黄图片识别数据集也各不相同【1】【2】。原因一方面在于，涉黄图片本身比较敏感，因法律法规问题通常不能公开在互联网中传播。另一方面，开源数据集对于涉黄图片的判别标准以及数据集本身的标注质量，数据质量都不尽如人意。这就使得不同识别方法之间不具有可比性。

如【图 1】所示，在对涉黄图片进行识别时，当前所有的公开方法中，都是将涉黄图片的识别问题当作是图片分类问题来处理（这点于违规图片识别的方法一致）。但要注意到，相对于一般的分类问题（如：ILSVRC 挑战赛中 1k 类的识别问题），涉黄图片分类时，类内之间样本的差异性要更为复杂。直接将一般的图片分类方法往涉黄图片识别领域迁移，能否得到好的结果，有带探究。

目前，国内外都已经出现比较成熟的商用的图片审核系统，如：腾讯云，网易云，阿里云上都有图片审核的 API 可供调用。开源的方法目前只看到有亚马逊开源的 open_nsfw【3】

涉黄图片判别标准

比较早的时候，特别是互联网还未普及时，图片涉黄的评价指标非常简单。只需要看图片中是否出现人的裸体就可以了【5】。随着互联网的普及，图片数量越来越多，种类也越来越丰富。简单的看图片中是否出现裸体这样的指标已经不能有效的解决图片涉黄鉴定的问题。

目前，找不到公开的，或者是一致的定量标准来判定图片是否涉黄。所有关于图片是否涉黄的判断都是定性的描述。所有的 paper 在谈到自己所用数据集的评价标准时，无一例外都是一笔带过，没有详细的解释，确实非常令人困扰。

先给出两个色情的定义，自行体会：

（1） Wiki 上对 Pornography 的定义

Pornography (often abbreviated porn) is the portrayal of sexual subject matter for the exclusive purpose of sexual arousal.

（2） 百度词典对“色情”的定义

色情（拼音：sè qíng），是一个汉语词语，是指能挑起或激发起性欲的东西。本指色欲、情欲。后引申为透过文字、视觉、语言描绘或表现裸体、性器官、性交等，与性有关的形象，使观赏者产生性兴趣和性兴奋的事物。

通过对当前公开的资料和论文整理，我认为图片涉黄判定标准可以归纳为以下几方面：

类别	评价标准	易错反例
性器官裸露	女性露出 3 点	艺术品，医用人体模型
	男性裸露 1 点	健美，健身
	儿童裸体	
	动物性交并裸露性器官	
性行为	单人/多人性行为	摔跤，沙滩，泳池场景
	同性亲密动作	
	动物性行为	
	性虐待	
敏感部位放大或特写	胸部/阴部/臀部放大特写	泳装，体育运动
	裙底偷拍或特写	
	动物性器官放大或特写	
性暗示	利用物品进行性挑逗	
	性诱惑动作	
	情趣服装	
	性动作模仿或者恶搞	
计生情趣用品	避孕套/震动棒/飞机杯等	
色情信息或者广告	招嫖等色情信息广告	
	性病治疗/药物广告	

表 1

从【表 1】中的信息可以看到，涉黄图片里包含的内容非常丰富。不同类别之间的差异性也非常大。一部分标准涉及人体器官，一部分标准涉及人的动作和行为，一部分标准涉及人与人之间的关系，还有涉及物品甚至于文本信息的。很显然，如果简单的将鉴黄理解成简单的 porn/nonporn 的二分类问题，显然不能很好的解决图片鉴黄的任务。

这几个标准中，关于性暗示的描述最为模糊。比如，如何定量“性诱惑”，“性挑逗”。这些往往是非常主观而且在不同国家地区差异性非常大的。因此，想要得到理想的图片涉黄识别数据集，需要训练一批经验丰富的数据标注人员。并且需要合理的设置正样本（SFW）。

涉黄图片识别开源项目和数据集

涉黄图片识别开源项目

目前，开源项目只有一个，来自 yahoo 的 open_nsfw【3】.

该项目的数据集没有公开。将涉黄图片识别看作是一个二分类的任务（NSFW / SFW）。仅仅给出一个预训练好的模型（目前，只开源了阉割版本 resnet50 这一个模型），提供给爱好者们去做实验。官方建议是在预训练好的模型的基础上做 finetune 实现不同场景下的应用。

结论，参考意义不大。

涉黄图片识别开源数据集

不论是出于版权的原因，还是出于政策法规的原因。当前，开源的涉黄图片识别开源数据集只有两个。一个是 Pornography Dataset【1】，公布于 2016 年；另一个数据集没有具体名称，来源于 GitHub 上的一个开源项目【2】，公布于 2018 年。

Pornography Dataset

这个数据集的作者来自 University of Campinas，Brazil（坎皮纳斯州立大学）。他们的研究组研究方向在于视频的理解。从 2014 年开始，他们有一系列的涉黄图像（包括图片和视频）识别的工作发表【1】。

Pornography dataset【1】来源于 Pornography-2k dataset【4】，它是后者的一个子集。Pornography dataset 包含了 800 个不同视频片段，总时长约 80h。通过对视频进行抽帧，最终得到数据集中的图片【图 2】。具体的抽取方法在“Data Processing”部分有具体介绍，可以理解为抽取视频关键帧。

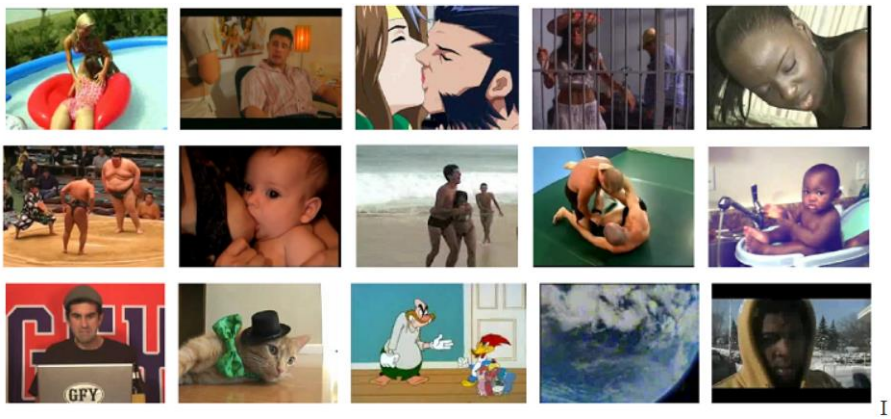


图 2 Pornography dataset 样本示例

800 个视频片段，“Porn/Non-Porn”各占一半。Porn 部分来源于成人电影网站，而 Non-Porn 部分来自一般网站。

“Non-Porn”又分成两部分，留意其中“difficult”部分。作者将"beach", "wrestling", "swimming"这 3 个不同场景下的视频视为 hard example。作者认为这些场景中出现的裸体，或者多人场景很容易被误认为是色情场景。

A summary of the Pornography database.

Class	Videos	Hours	Shots per Video
Porn	400	57	15.6
Non-porn ("easy")	200	11.5	33.8
Non-porn ("difficult")	200	8.5	17.5
All videos	**800**	**77**	**20.6**

图 3 Pornography dataset 数据构成（1）

在挑选“Porn”这一类型的视频时，作者有意识的选择不同肤色种族的人，以便数据集更具有代表性。

Ethnic diversity on the pornographic videos.

Ethnicity	% of Videos
Asians	16%
Blacks	14%
Whites	46%
Multi-ethnic	24%

图 4 Pornography dataset 数据构成（2）

这个数据集最大的问题在于，它所制定的涉黄于非涉黄的标准非常粗糙。它将 Porn 视频片段中所有的视频帧都标注为“Porn”。很显然，对于带剧情的成人电影来说，并非所有视频帧都涉黄。而且，从图片来源上说，仅仅从成人电影中抽帧得到的图片，其本身的多样性就非常小。想要泛化到其他场景，堪忧。

这个数据集可以免费下载，但是需要签一个协议并且邮件作者来获取真正的数据地址。并且作者回复邮件速度非常快 lol。

Github Project：nsfw_data_scrapper

这批数据来源于公开的图片网站（Ripme）以及成人网站。项目只提供图片 url 而不提供图片本身。

数据集一共划分为 5 个不同的类别，每个类别的判别标准如下：

- porn - pornography images
- hentai - hentai images, but also includes pornographic drawings
- sexy - sexually explicit images, but not pornography. Think nude photos, playboy, bikini, etc.
- neutral - safe for work neutral images of everyday things and people
- drawings - safe for work drawings (including anime)

作者在项目开始就有强调：“Disclaimer: the data is noisy - do not use to train a production model unless you want negative media coverage!”

这里透露出几个信息：（1）数据没有经过人工标注。（2）每个类别的图片多半是通过关键字检索得到的，精度不高。

因为数据集质量问题，后面给出的所谓 acc 这些指标其实没有多大的参考价值。而且，从【图 5】可以看出，porn 这个类别的图片数量要远远大于其他类别，类别间样本不均衡问题也比较显著。

```
Number of URLs in ../raw_data/drawings/urls_drawings.txt:
   25732
Number of URLs in ../raw_data/hentai/urls_hentai.txt:
   45228
Number of URLs in ../raw_data/neutral/urls_neutral.txt:
   20960
Number of URLs in ../raw_data/sexy/urls_sexy.txt:
   19554
Number of URLs in ../raw_data/porn/urls_porn.txt:
  116521
```

图 5 每个类别图片数量

涉黄图片识别方法

涉黄图片识别的文章最早可以追溯到 20 世纪 90 年代【5】。涉黄图片的识别作为图像识别的一个子集，识别涉黄图片的方法伴随着识别图像的方法而逐渐发展起来。可以看到，当一般图像识别领域出现新的方法时，这个方法马上会被应用到涉黄图片的检测上，并且效果上似乎都有所提升。目前，涉黄图片识别的方法的综述我比较推荐【6】.

涉黄图片识别方法的发展经历了以下 3 个阶段：

时间	2000年以前	2000-2012	2012之后
视觉特征	肤色/裸体	肤色/三点/肢干/动作	综合
抽象特征	颜色+纹理	颜色+纹理+SIFT → VBOW/VLAN	综合
分类器	RL / k-nearest /SVM	SVM	CNN

图 6 涉黄图片识别方法的发展

我们可以同时对比图像识别领域的方法演进（【图 7】修改自【7】）：

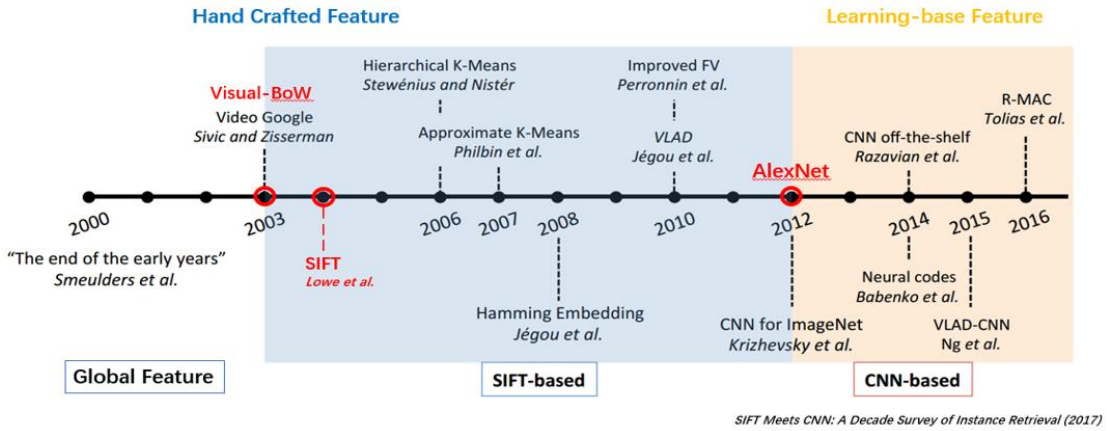
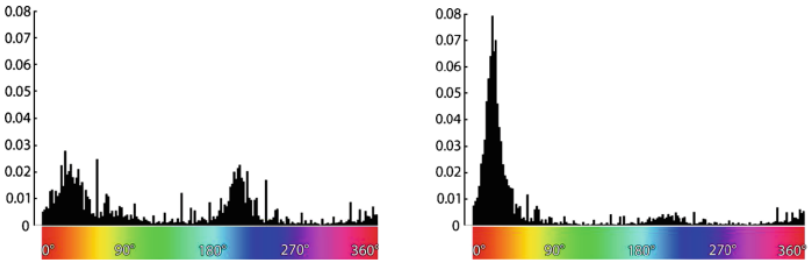


图 7 图像识别领域的方法演进

最早从事涉黄图片识别的工作，就是判断图片中是否出现人的裸体，而人的裸体最显著的特征就是图片色调偏暖（应该忽略了肤色问题）【图 8】【6】，而且人体肤色占据图片大部分。所以，最早是以颜色特征作为识别涉黄图片特征【5】。后来，发现光颜色这一特征不够稳定，在颜色的基础上增加纹理特征来识别图片中是否出现人的躯体。通过识别颜色和人体两个维度判断图片是否涉黄。同时期在图像识别领域，识别图像基本上也是依靠如颜色，纹理这样的全局特征来对图像进行特征提取，载构造分类器对图片分类。



(a) Flickr (b) Adult

Fig. 1 The hue value distribution of the pixels of **a** 100,000 random images downloaded from Flickr and **b** 100,000 images with adult content (including background pixels). The hue value is represented by an angle between 0° and 360° according to its definition in the HSV color space. The ratio of pixels in the skin color range (the *red-orange* area roughly with an extent of 30°) is notably higher for adult images

图 8

Flickr 所有图片统计颜色直方图（左）

Porn 图片统计颜色直方图（右）

2003 年出现的 BOVW（Bag of Visual Word）【9】以及 2004 年出现 SIFT【8】使得图像的 handcrafted 特征由全局特征向局部特征演进。通过图像局部特征构建更为抽象的全图 BOVW 或者 VLAN 特征，再结合 SVM 分类器，成为这个时期图像识别领域的标准方法。这个时期的涉黄图像识别也是这个思路，具体的 pipeline 如下【图 9】：

图 9 涉黄图像识别 pipeline

不同的工作在于不同的提特征的方法和分类器之间的组合。看哪种组合能够获得更好的结果。

2012 年大放异彩的 CNN 方法也同样惠及这个领域。2012 年以后，基本上所有涉及涉黄图像识别的工作都在使用 CNN 方法来实现。

1996-Finding Naked People【5】

将涉黄图片检测分成两个步骤实现【图 10】：（1）利用颜色信息，将图像中表示人体肤色的像素分割出来；（2）对于（1）中分割出来的部分，利用纹理信息构建人体躯干。利用人体躯干的拓扑结构识别分割部分像素是否是人体。是，则图片是涉黄图片；否则，是正常图片。

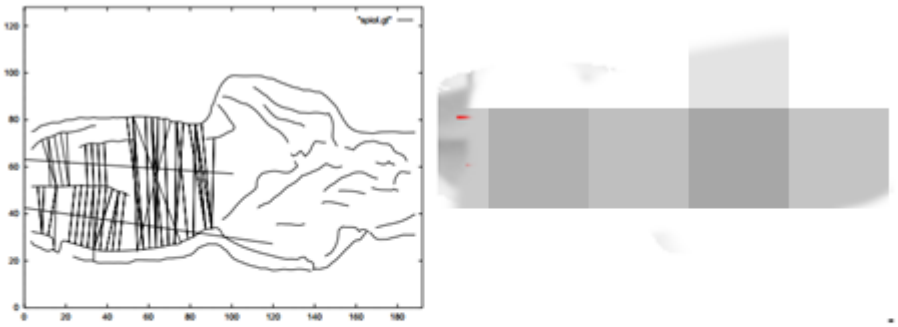


Fig. 2. Grouping a spine and two thighs: *Top left* the segment axes that will be grouped into a spine-thigh group, overlaid on the edges, showing the upper bounds on segment length and the their associated symmetries; *Top right* the spine and thigh group assembled from these segments, overlaid on the image.

图 10

2006-Large scale image-based adult-content filtering【10】

Google 出品，一般来说必属精品。采用了标准的涉黄图片识别 pipeline【图 9】来实现。

Feature Extractor 部分：

特征属性	特征数量
SKIN COLOR	2
CONNECTED COMPONENT ANALYSIS	4
SKIN TEXTURE FEATURES	2
LINES	1
IMAGE FEATURES	3
ENTROPY FEATURES	2
CLUTTER FEATURES	2
FACE DETECTION	2

总共提取了 18 个特征作为一张图片的特征。

Classifier 部分：

使用标准 LIBSVM 库搭建分类器，实现 porn/nonporn 的二分类。

2008-Bag-of-Visual-Words Models for Adult Image Classification and Filtering【11】

为数不多的，在文章中有明确给出数据集标注标准【图 11】。可以看出，这个数据集以图中是否出现裸体为一个关键的指标给与标注。现在看来，这样的标注显然不能 cover 复杂的涉黄图片识别场景，但标准至少给出来了。

class 0: inoffensive images,
class 1: lightly dressed persons, might be offensive in very strict environments,
class 2: partly nude persons, might be objectionable in school environments,
class 3: nude persons, likely objectionable in many environments, and
class 4: porn images, probably offensive in most environments.

图 11 数据集类别和每个类别标注标准

Feature Extractor 部分：

DOG + random patch + PCA 来代替 SIFT(sift 用灰度图，丢失颜色信息，而作者认为颜色信息对鉴黄极度重要，不然干嘛把“nude person”作为标注指标)。构建 BOVW 特征。

Classifier 部分：

SVM 或者 log-linear models 实现 1 vs. rest 的分类。

2015-A Comparative Study of Local Feature Extraction Algorithms for Web Pornographic Image Recognition

这篇文章的方法也是做鉴黄的一种场景思路。利用 Coarse-to-fine 的思想实现图片鉴黄【图 12】。Coarse 部分做人脸和肤色检测，过滤很大一部分图片。Fine 部分，利用图像 local feature 构建 BOVW 或 VLAN，结合 SVM 分类器最终实现涉黄图片检测【图 13】。

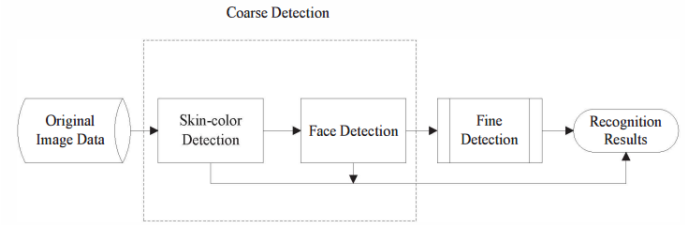


Fig.1. The proposed scheme in this paper

图 12

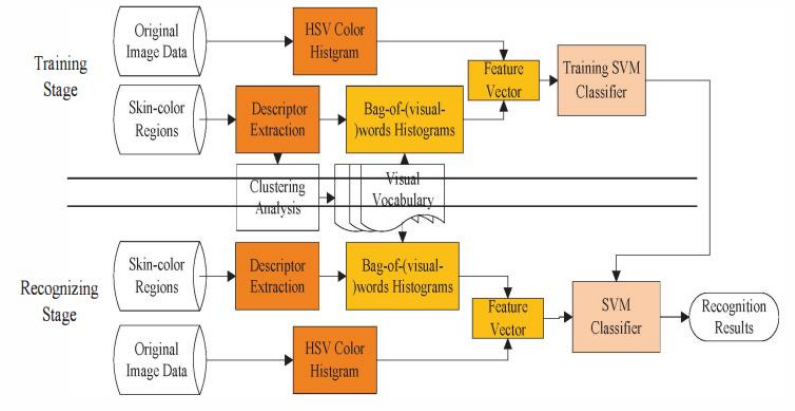


图 13

2015-Applying deep learning to classify pornographic images and videos【13】

使用 CNN 做图片鉴黄最早的 paper 之一。只做 porn/nonporn 的二分类。数据集使用的是开源数据集【1】。

具体实现上，将 imagenet 数据集预训练好的 AlexNet 和 GoogleNet 直接在自己的数据集上做 finetune。测试的时候，将两个模型的输出结果做投票得到最终的分类结果【图 14】。

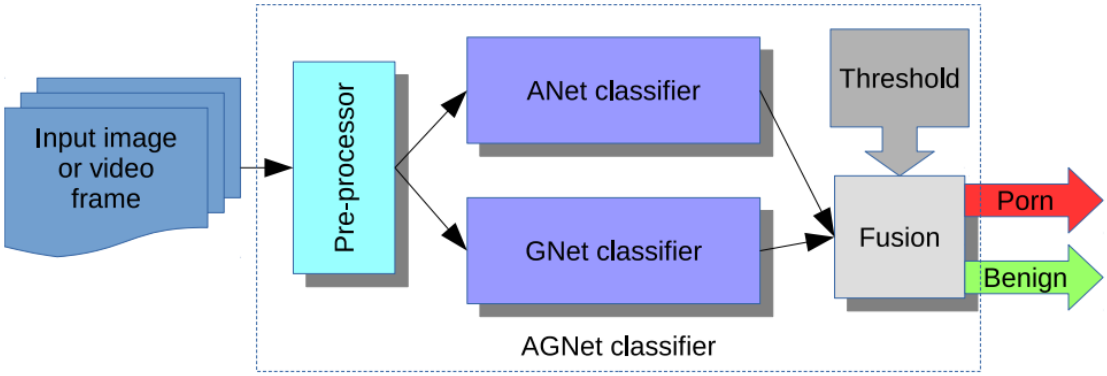


Fig. 3. Proposed AGNet porn image classifier.

图 14 AGNet 结构

如【图 14】，所谓 ANet 就是 AlexNet，而 GNet 就是 GoogleNet。这两个模型分别 finetune。

2016-Bootstapping Deep Feature Hierarchy For Pornographic Image Recognition【14】

这篇文章来自中科院，也是基于 CNN 实现图片鉴黄。只不过思路相对一般的 CNN 方法路子比较野。作者认为像肤色，纹理这样的 low lever 特征对于图片鉴黄很重要，需要增加权重。所以作者将 CNN 不同层之间的 feature 通过 SPP-layer 做了 resize 然后在组合起来【图 15】，作为新的特征训练分类器【图 16】。

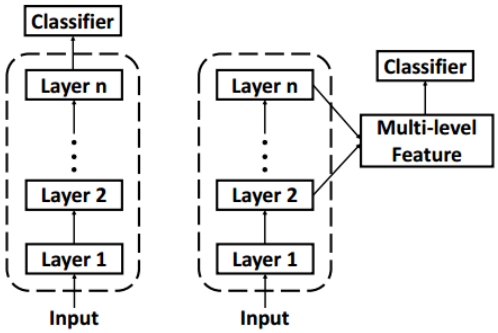
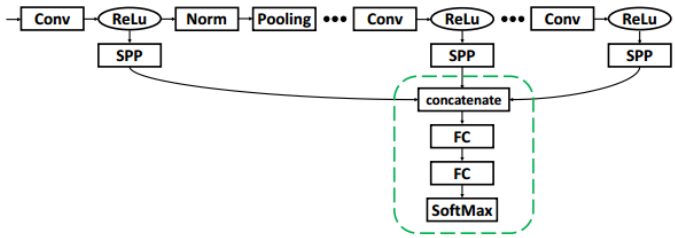
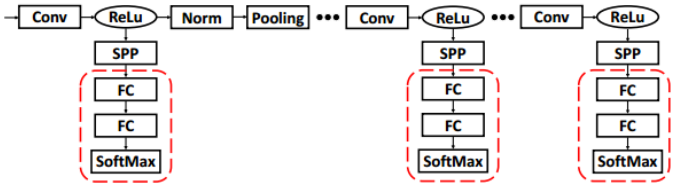


Fig. 1. Schematic diagrams of mainstream CNNs for image classification (**left**), and our proposed MLFF-CNN for pornographic image recognition (**right**). Our architecture incorporates features from multiple levels to perform recognition.

图 15



(a)



(b)

Fig. 2. The detailed architectures of our MLFF-CNN (**a**), and PT-CNN which is used for pre-training (**b**).

图 16

而且在训练的时候，增加了在线的 hard negative mining，增强模型的训练。

可以从【图 17】看到，融合的中间层越多，识别的效果越好；

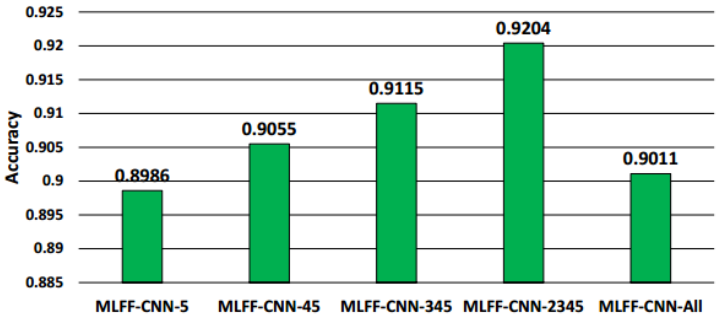


Fig. 3. Comparison between five networks which increasingly incorporate more layers.

图 17

2017-Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-Coarse Strategy【15】

依旧是中科院出品，有两点需要特别关注：（1）将物体检测思路应用到涉黄图像识别上；（2）支撑（1）的前提条件是，利用他们内部数据集“Sensitive Dataset”，对图片中涉黄部位做 bounding box 的标注【图 18】。

Fig. 3. Some confusion examples. The first row illustrates the full images, and images in the second row are local regions with respect to the full images. The first and second columns are normal images, whereas the third and fourth columns should be considered as adult images.

图 18 Sensitive Dataset 示例

网络结果如【图 19】，最底层是一般的分类网络结构；最上层是经典的物体检测网络 Faster-RCNN，中间部分有点 SSD 的影子在。总之，就是结合 local 和 global 的信息，实现图片鉴黄。模型最终输出 3 个类别：

（1） **normal**（995 classes are defined by ImageNet, and two categories are defined by our collections）

（2） **adult**（nine remarkable eroticism and nudity fine-grain categories e.g., nudity, oral sex, sexual organs, sexual behaviors）

（3） **unsuitable for children**（e.g., underwear, swimwear, leggy model）

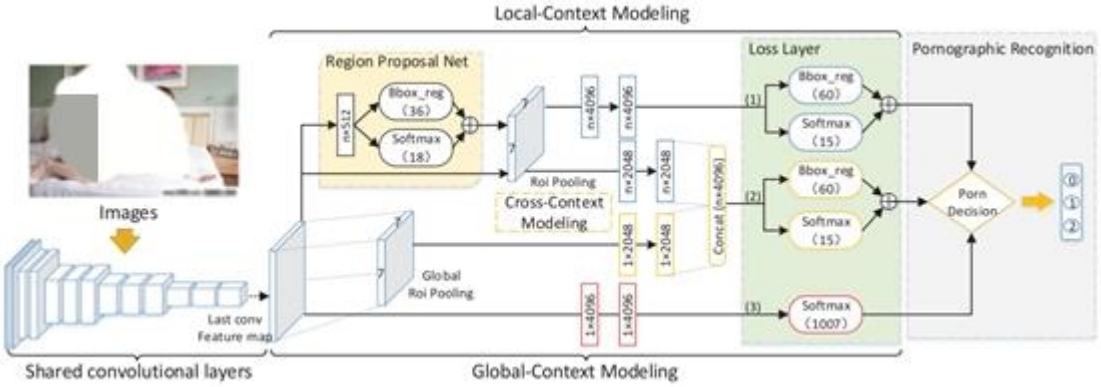


图 19

所谓“fine-to-coarse”就体现在 loss layer 到最终输出的这部分。网络的每一个 branch 都是单独训练的，先训练 global context 部分获得 base model。然后再训练其他两个 branch。训练 local 和 cross-context 这两个 branch，只针对一部分类别（性器官，敏感部位）做 detection 的训练。

最终输出类似一个一层的 maxout 的网络【图 20】，直接将 3 个 branch 的输出做了融合，完成“fine-to-coarse”操作：

$$\mathcal{F}_{DMCN} = \max((1-w_1-w_2)\cdot\tilde{\mathcal{F}}_{global}, w_1\cdot\tilde{\mathcal{F}}_{local}, \quad w_2\cdot\tilde{\mathcal{F}}_{cross}). \tag{6}$$

图 20

这个文章，其实相当于用了多模型融合实现涉黄图片识别。与单模型方法【13，14】相比感觉优势不是特别大。而且，还需要做专门的数据标注工作。但它提醒了一点，attention 机制如果使用在黄图鉴别上或许能提升鉴别的效果。

挑战

图片鉴黄最大的挑战还是在于数据集和数据集评价指标的统一问题。到目前为止，各个不同的方法比较来比较去，都是在不同的数据集上进行，几乎不具备可比性。而且，不同研究者之间的数据也不公开。所以，在这个方向想要找到或者产出特别好的 paper 挺不容易。

涉黄类型的图片多种多样，想要将不同类型的图片统一合并成一个类别给与“Porn”或者“NSFW”的标签，然后做分类。恐怕是不能很好的解决涉黄图片识别的问题的。比如，有一些文章就单独对涉黄图片的一个子类做研究【16】（仅对 under skirt 裙底偷拍场景做研究）。

除了图片以外，文本以及视频涉黄的鉴别目前也是非常热门的研究方向。

参考文献

1. Pornography Dataset：https://sites.google.com/site/pornographydatabase/
2. Github project “nsfw_data_scrapper”：https://github.com/alexkimxyz/nsfw_data_scraper
3. Yahoo Open NSFW project: https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for
4. Moreira, D et.al., “Pornography classification: The hidden clues in video space-time”. (2016)
5. Margaret Fleck, “Finding Naked People”. (1996)
6. Ries Christian X et.al., “A survey on visual adult image recognition”. (2012)
7. Liang Zheng et.al., “SIFT Meets CNN-A decade survey of instance retrieval”, (2017）
8. D. G. Lowe et.al., “Distinctive image features from scale-invariant keypoints”, (2004)
9. J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos” , (2003)
10. Henry A. Rowley et.al., “Large scale image-based adult-content filtering”，（2006）
11. Thomas Deselaers et.al., “Bag-of-Visual-Words Models for Adult Image Classification and Filtering”，（2008）
12. Zhen Geng et.al.，“A Comparative Study of Local Feature Extraction Algorithms for Web Pornographic Image Recognition”，（2015）
13. Mohamed N. Moustafa et.al.，“Applying deep learning to classify pornographic images and videos”，（2015）
14. Kai Li et.al.，“Bootstapping Deep Feature Hierarchy For Pornographic Image Recognition”，（2016）
15. Ou, Xinyu et.al.，“Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-Coarse Strategy”，（2017）

Yi Huang et.al.,“Using a CNN Ensemble for Detecting Pornographic and Upskirt Images”,（2016）