第五章 网络多媒体信息内容安全

5.1 概述

5.2 网络不良图像内容识别

5.3 网络不良视频内容识别

5.4 网络不良音频内容识别

5.1 概述

- 随着多媒体通信技术的迅速发展，大量的文本、语音、视频、图片等多媒体信息成为了互联网的主要信息元素，匿名加密多媒体等业务也广泛使用。
- P2P技术和安全加密技术传送多媒体信息的广泛应用，使得通信变得越来越安全和高效。

5.1 概述

然而这些多媒体信息流也成为了大量反动、色情、无用垃圾信息的“传送带”，侵占了网络带宽等资源，从而破坏了互联网健康有序的绿色环境。为营造健康、和谐与稳定的互联网环境，必须实现对特定多媒体信息流的有效、灵活、可扩展地识别和过滤。

5.1 概述

目前对互联网不良多媒体信息的过滤方法：

1. **基于分级标注的过滤**
2. **基于URL的过滤**
3. **基于关键词的过滤**
4. **基于内容分析的过滤**

1. 基于分级标注的过滤

- 使用浏览器本身或第三方特别是PICS（Platform for Internet Content Selection）和ICRA（Internet Content Rating Association）分级标注过滤。
- 具体是用户或管理员通过浏览器的安全设置选项实现网页内容过滤。
- 并不是所有的网站都遵守ICRA标准，使得基于分级标注的过滤成为形同虚设

2.基于URL的过滤

- 将已知有害页面和网站收集到URL禁止列表库，将允许访问的网页和网站收集到URL允许列表库，即设置网页黑白名单。过滤系统检测到某网络地址在黑名单中时，将过滤该网络地址以阻止用户访问，否则，将放行。

优点：简单易实现。

缺点：

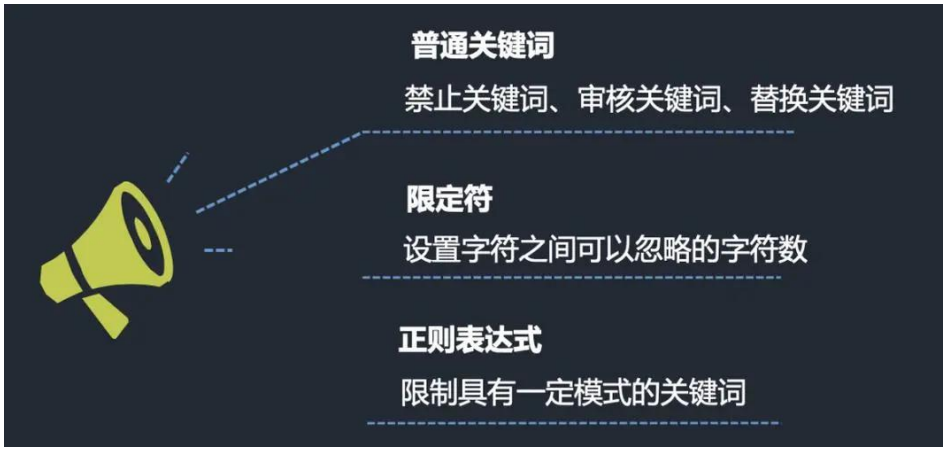
不够灵活

管理难度大

数据库维护难度大

3.基于关键词的过滤

- 对文本内容、文档的元数据、检索词、URL等进行关键词匹配，再对满足匹配条件的网页或网站进行过滤。
- 具体就是从网页中提取出关键词与预先建立的不良或敏感关键词数据库匹配，通过设定阈值计算匹配程度来判断是否为不良网站，是则过滤该网站，否则放行该网站。



优点：简单易实现。

缺点：

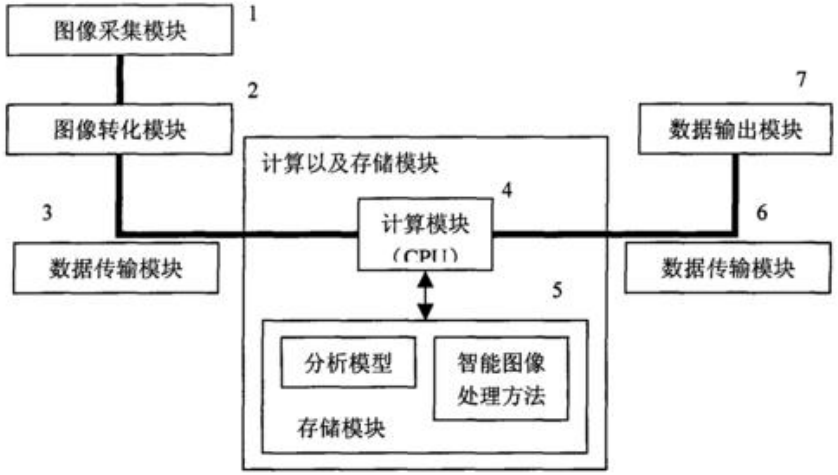
错误率较高

过滤范围窄

不够灵活

4.基于内容分析的过滤

■通过语义分析、机器学习、图像处理等技术分析用户浏览的网页内容来判断该网页是否该过滤 。

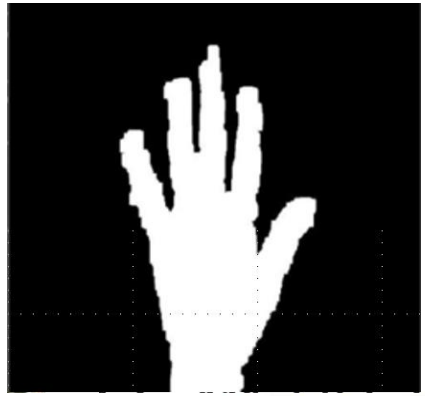


优点：过滤准确度高

缺点：技术难点大、过滤速度慢

5.2 网络不良图像内容识别

- 在不良多媒体信息中，不良图像是色情信息的重要载体。
- 图像内容过滤技术是根据图像的色彩、纹理、形状、轮廓以及它们之间的空间关系等外观特征和语义作为索引，通过与人体敏感部位相关数据进行相似度匹配而进行的过滤技术。



5.2 网络不良图像内容识别

根据不良图像自身的特点，一般从三个角度来进行不良图像的判定识别：

1. 从皮肤裸露情况来判断
2. 从敏感部位来判断
3. 从猥亵的人体姿态来判断

1.从皮肤和敏感部位裸露情况来判断

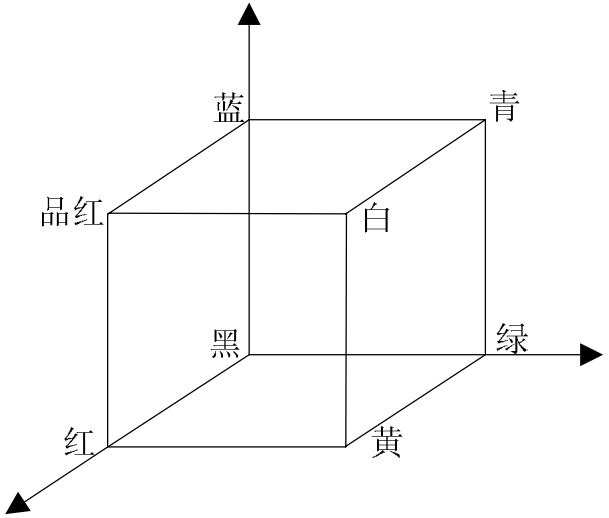
- 对于皮肤裸露面积较大的情况，可以先从图像的低层特征出发，找出多个可以较好区分不良图像和正常图像的特征，再组合这些特征应用于智能分类器实现不良图像分类。这一实现方法思路简单，可行性强，是目前研究者们的首选。
- 对于敏感部位裸露的情况，可以为敏感部位建模，以此建立起分类器。这一方法较为接近人类理解敏感图像的方式，是过滤不良图像的最直接途径，但由于不良图像过于复杂，这些特殊部位的定位和模型建立又很困难，因此实现起来难度很大。

2.从猥亵的人体姿态来判断

- 此种情况可以尝试识别出图像中的人体肢体，再根据模板或内建的规则进行组建，如果组建成功，则该图像为不良图像，否则为非敏感图像。
- 困难
  - 在复杂的图像背景下如何正确识别出人体肢体
  - 适当的模板和组建规则的确定
- 由于不良图像本身姿态的不确定性，目前通过这种方法进行人体检测的系统相对较少。

5.2 肤色检测

- 肤色检测是在图像中选取对应于人体皮肤像素的过程，通常包括颜色空间变换和肤色建模两个步骤。
- 1.颜色空间：颜色的描述是通过颜色空间来实现的，不同的颜色空间应用于不同目的和处理情况。颜色空间的分类有很多，常见的有RGB、YUV、HSV、HIS等。



1.颜色空间

RGB到YCbCr的转换如下所示：

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix}$$

从上式可以看出，YCbCr颜色空间是由RGB线性导出的颜色空间，类似的线性变换颜色空间还有OPP、YIQ、YUV等颜色空间。

1.颜色空间

RGB到HSV的转换如下所示：

$$\begin{cases} H = \arccos \dfrac{\frac{1}{2}\left[(R-G)+(R-B)\right]}{\sqrt{(R-G)^2+(R-B)(G-B)}} \\ S = \dfrac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)} \\ V = \dfrac{\max(R,G,B)}{255} \end{cases}$$

从上式可以看出，HSV颜色空间是由RGB**非线性**导出的颜色空间，类似的线性变换颜色空间还有YIQ、YUV等颜色空间。

2.肤色模型

- 肤色模型是关于肤色知识的计算机表示，通过训练样本集建立肤色模型是肤色检测的关键，常用的三种肤色建模方法是：
  - 肤色区域模型
  - 高斯分布模型
  - 统计直方图模型

2.肤色模型

肤色区域模型：符合公式中的像素认为是皮肤像素

- RGB颜色空间：

$$\begin{cases} R>95, \text{且} G>40, B>20 \\ \max(R,G,B)-\min(R,G,B)>15 \\ |R-G|>15 \text{且} R>G, R>B \end{cases}$$

- YCbCr颜色空间：

$$\begin{cases} 77 \le Cb \le 127 \\ 133 \le Cr \le 173 \end{cases}$$

- HSV颜色空间：

$$\begin{cases} 0 \le H \le 0.1388 \\ 0.23 \le S \le 0.68 \\ 0.35 \le V \le 1 \end{cases}$$

2.肤色模型

高斯分布模型：单高斯模型和高斯混合模型

- 单高斯模型：计算像素x属于肤色的概率

$$p(x|skin)=\frac{1}{2\pi|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

- 高斯混合模型：计算$p(x,\mu,\Sigma)$ 是否大于阈值

$$p(x,\mu,\Sigma)=\sum_{i=1}^{M}\omega_i\frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right]$$

2.肤色模型

统计直方图模型：通过肤色样本的直方图统计构造肤色概率图SPM（Skin Probability Map）进行皮肤检测

- 规则化查找表：将大于下式的像素认为是肤色

$$p_{skin}(x) = \frac{count(x)}{Norm}$$

- 贝叶斯分类器：

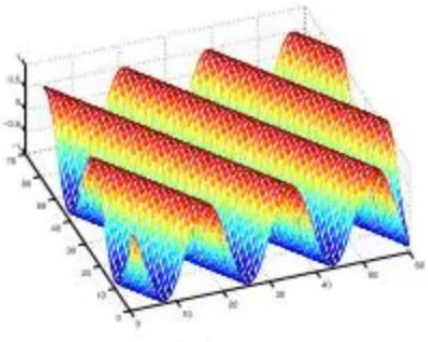
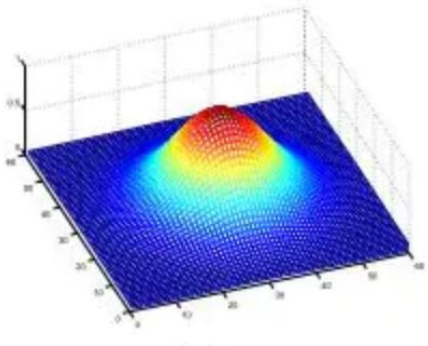
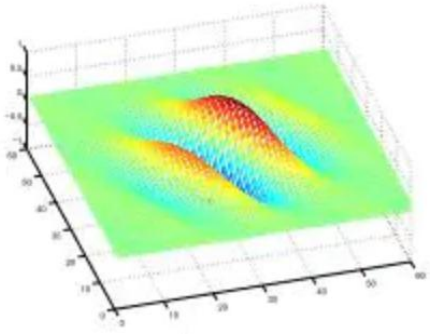
$$p(skin|x) = \frac{p(x|skin)\,p(skin)}{p(x|skin)\,p(skin) + p(x|-skin)\,p(-skin)}$$

5.2.2 纹理分析

对皮肤纹理进行分析的目的在于提高对不良图像皮肤区域检测的准确性。人体皮肤纹理作为一种特殊的纹理，没有明显的纹理基元，没有明显的周期性和方向性，它的一个重要特征是光滑，所以皮肤区域中的灰度变化比较小，区域灰度方差值也比较小，反映在灰度上就是图像的灰度包络平滑，变化缓慢，而非皮肤图像一般没有这个特征。

1. Gabor滤波法

- Gabor滤波器纹理分析方法就是选用某一特定的Gabor函数，然后设计一种Gabor滤波器，用设计好的Gabor滤波器去过滤图像，对过滤后的图像再提取能量统计特征作为纹理特征。
- Gabor 变换是短时傅里叶变换中窗函数取为高斯函数时的一种特殊情况。因此，Gabor滤波器可以在频域上不同尺度、不同方向上提取相关的特征。另外，Gabor函数与人眼的作用相仿，所以经常用作纹理识别上，并取得了较好的效果。

2. 灰度共生矩阵法

灰度共生矩阵是对图像上保持一定距离的两像素分别具有某灰度的状况进行统计得到的，描述了成对像素的灰度组合分布，可以看成是两个灰度组合的联合直方图。灰度共生矩阵反映了纹理关于方向、相邻间隔、变化幅度的综合信息，既反映纹理的粗糙程度，也反映纹理的方向性。

图像纹理特征中的四种最常用特征：

- 角二阶矩：$ASM = \sum_{h}\sum_{k}(m_{hk})^2$
- 熵：$ENT = -\sum_{h}\sum_{k} m_{hk}\log m_{hk}$
- 对比度：$CON = \sum_{h}\sum_{k}|h-k|\,m_{hk}$
- 相关性：$COR = \left[\sum_{h}\sum_{k} hkm_{hk} - u_x u_y\right] \Big/ \sigma_x\sigma_y$

3. 灰度统计法

灰度统计法是数字图像处理的基本方法。

步骤如下：

先对收集的皮肤区域像素进行灰度统计，得到皮肤区域的统计信息，作为皮肤纹理特征的经验值；

当进行皮肤纹理提取时，将待检测区域像素同经验值相比较，满足一定关系时，判定为皮肤纹理。

5.2.4 不良图像的识别

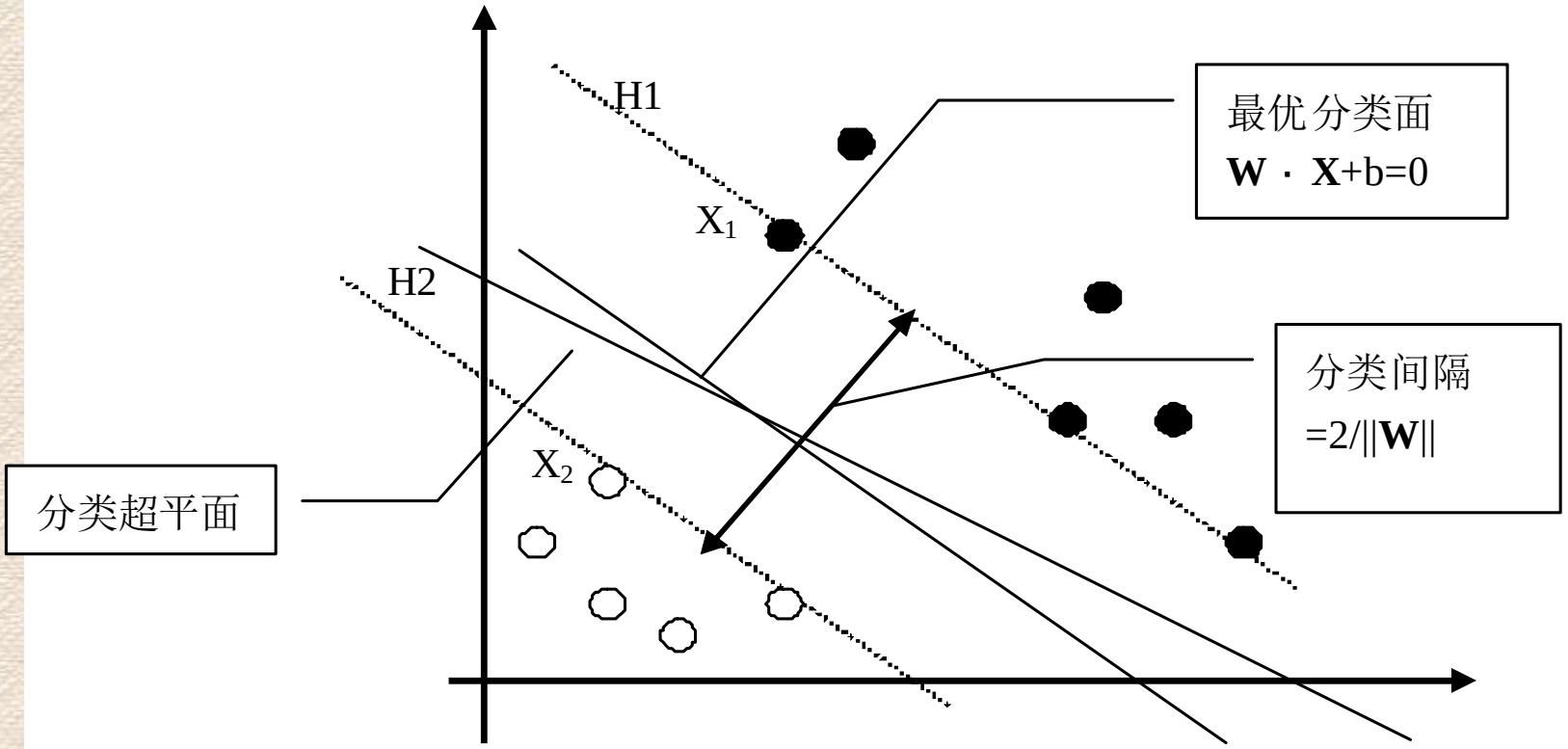
- 经验阈值法是最早出现且最简单的不良图像识别方法，它利用皮肤颜色分布聚合在一块压缩区域特点，从色彩空间上规定阈值来区分出皮肤。
- 该方法过度依赖样本和样本空间选择，且阈值选择缺乏一定适应性。
- 支持向量机分类器已成功应用于图像识别，并以明显优势取得了比以往分类器更好的效果。

1. 支持向量机

- 支持向量机（Support Vector Machines, SVM）： 20世纪90年代初由Vapnik提出的一种机器学习方法。

- SVM的构造是通过在特征空间中构造最优分类超平面（Optimal Hyperplane），最优分类超平面是指两类的分类空隙最大，即每类距离超平面最近的样本到超平面的距离之和最大。距离这个最优超平面最近的样本被称为支持向量（Support Vector）。

1. 支持向量机



最优分类面

1. 支持向量机

在线性可分的情况下，存在多个超平面(Hyperplane)如H1、H2等，使得这两类被无误差的完全分开，这个最优分类超平面被定义为:

$$W \bullet X + b = 0, W \in R^n, b \in R$$

其中 $W \bullet X$

是内积（dot product），*b*是标量。

1. 支持向量机

SVM在应用过程中只做训练样本之间的内积运算，这种内积运算由事先定义的核函数实现，与核函数的结合，使支持向量机的适用范围更广。

常用核函数：

- 线性函数：　　如 $K(x,y)=x^r y$
- 多项式函数：　　如 $K(x,y)=\left(\lambda x^r y+r\right)^d, \lambda>0$
- 径向基函数：　　如 $K(x,y)=\exp\left(-\|x-y\|^2 \Big/ \sigma^2\right)$
- Sigmoid内积函数：如 $K(x,y)=\tanh\left(\lambda x^r y+r\right)$

2. 基于SVM的不良图像识别

不良图像识别问题是一个两类分类问题，其判别函数为：

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K(x,x_i)+b\right)$$

SVM训练算法的本质是求解一个二次规划问题，即最优化该问题的解就是要使得所有样本都满足如下条件(Kuhn-Tucker条件)：

$$\alpha_i=0\Leftrightarrow f(x_i)\geq 1$$

$$0<\alpha_i<C\Leftrightarrow f(x_i)=1$$

$$\alpha_i=C\Leftrightarrow f(x_i)\leq 1$$

其中C为用户定义的常量，用以表示模型复杂度与分类错误率之间的一种平衡，$f(x_i)$为SVM相对于第i个样本的输出。

2. 基于SVM的不良图像识别

如果样本规模过大，则有可能使得矩阵

$D = y_i y_j K(x_i, x_j)$ 过大进而使得无法用计算机来完成处理工作。于是如何使得SVM对大规模样本集的训练能力得以提高与如何精简样本集来提高SVM的训练速度成为SVM研究领域中的热点问题。

5.3 网络不良视频内容识别

根据网络视频所使用的网络传输模式，可将网络视频基本上划分为两大类，

- 基于传统的B/S模型的在线网络视频，如优酷网、土豆网、酷6网、YouTube等网站上的视频，称为Ⅰ类型网络视频；
- 另一种是基于P2P网络的网络视频，如PPlive、UUSee、TVKoo等软件产生的网络视频，称为Ⅱ类型网络视频。
- 不同传输模式的视频的监管不同。

5.3.1 网络视频流的发现

对网络视频数据流的发现首先是识别应用层协议。其次，应当对应用层协议内容进一步判断，这就需要对视频流进行特征分析。此外，由于网络视频流存在着多种不同类型，还应当将不同类型的网络视频流进行区分，以方便视频内容监管等后续的处理。

1 应用层协议识别

- 基于特征串频率的识别方法
  - 以协议中出现频率最高的字段作为特征串来识别协议的方法，采用一个特征串来标识一种协议
- 基于签名字串的识别方法
  - 针对P2P协议的范围，需要对整个报文通过匹配多个特征串来识别一种P2P协议。
- 其他方法等。

2.协议内容的识别

网络视频流交互过程中的特征：

- 不同的网络视频流，其交互过程各阶段使用的协议可能相同，也可能不相同；
- 不同关键特征字串(CRS)对应着不同类型的网络视频流，一种CRS可标识一种类型的网络视频流；
- CRS在数据包中的位置具有稳定性；
- 使用相同流媒体信令协议(SMSP)的不同网络视频流，其所包含的关键字段特征参数(CP)是不同的。

5.3.2 网络视频流流量的获取

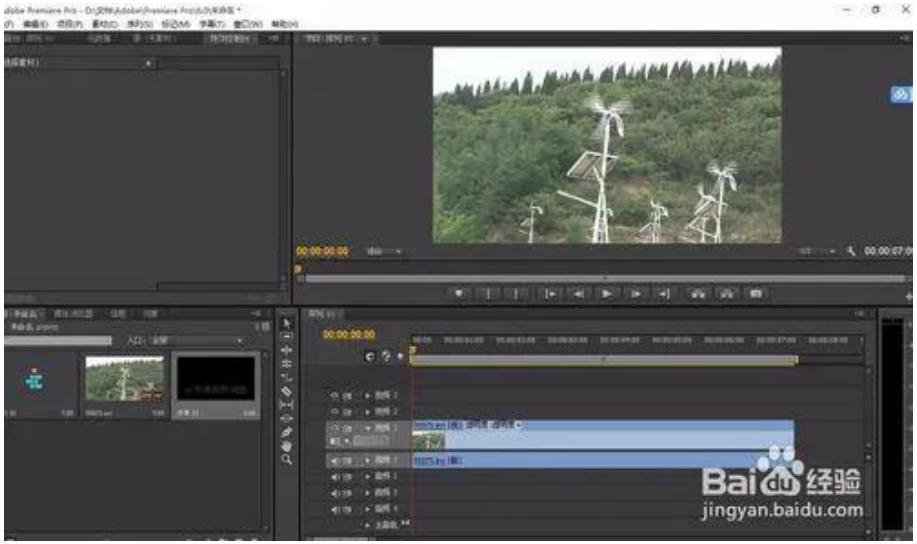
- 基于端口的流量识别方法
  - 依据互联网地址指派机构IANA规定的端口映射表，通过截取数据包头的端口信息，如果可以匹配某个已知端口，可以直接识别流量。（23、80）
- 基于净荷特征的流量识别方法
  - 找出交互过程中不同于其它协议的字段，作为协议的特征，通常采用DPI来匹配各种应用特征，从而识别出不同的多媒体流量。
- 基于统计行为特征的流量识别方法。
  - 依据机器学习及数据挖掘的统计决策、聚类等模式分类思想。

5.3.3 视频时域分割

- 视频是有结构层次的，这种层次体现在分段管理和帧之间的时间顺序上。
- 视频编辑主要以镜头为单位，进行时域分割时必须研究镜头间的过渡关系、转换方式，主要对象包括镜头切变和渐变。
  - 切变：前景、背景不同，视觉特征（颜色、区域形状、纹理等）也存在突变。
  - 渐变：以逐步替换的划变和淡入淡出的溶解二者为代表。

5.3.4 视频关键帧提取

关键帧是用于描述一个镜头的关键图像帧，连续的关键帧序列通常反应了视频的主要内容。关键帧的选取是在视频中各个镜头内挑选出具有代表性的静态图像，作为视频内容分析的主要对象，它既是前面视频时域分割的目的之一，也是连接静态图片识别处理方法的桥梁，具有重大意义。



5.3.4 视频关键帧提取

关键帧的选取至少需符合两个基本条件：代表性和简单性。代表性要求反映视频主要内容，保有可复现的重要细节；简单性要求信息冗余小，能降低检测计算的复杂度。

常用提取方法：

- 基于固定帧采样的提取方法；
- 基于帧间差的关键帧提取方法；
- 基于聚类的关键帧提取方法；

5.3.4 视频关键帧提取

- 基于视频单元分类的关键帧提取方法；
- 基于累积帧间差的关键帧提取方法；
- 基于运动信息的关键帧提取方法；
- 基于文字和图像信息的关键帧提取方法；
- 基于MPEG压缩流的宏块统计特性提取关键帧

5.4 网络不良音频内容识别

视频流中的音频信号是一种或多种声音信号（语音、音乐以及噪声等等）交织在一起的复杂混合体。对音频信号分析的目的是能够对音频信号进行分类，把不同类别的声音信号区分开来。网络不良音频内容识别首先需要对音频进行特征分析，然后应用音频分类技术对音频内容进行分类识别。

5.4.1 音频数据预处理

对音频进行处理之前，通常要进行预处理，将音频流切分成时间长度较短的单元，音频数据预处理模块主要实现以下两个功能：

- 对原始音频数据做预加重；
- 对预加重之后的信号进行加窗分帧，形成音频帧，为音频信号特征提取做准备 。

5.4.1 音频数据预处理

将音频信号分成一些段时间段来处理，在这些段中具有固定的特征，这种分析处理方法称为“短时”分析方法。从音频流中切取出短时音频段的过程称为分帧。分帧的方法一般如下：用一个长度有限的窗序列来截取一段声音信号进行分析，并让这个窗滑动以便分析在任意时刻附近的信号。

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)] \times \omega(n-m)$$

其中，$T[]$ 表示某种运算；$\{x(m)\}$ 为输入信号序列。

5.4.1 音频数据预处理

最常用的三种窗函数：

- 矩形窗：

$$\omega(n)=\begin{cases}1, & 0\le n\le N-1\\ 0, & otherwise\end{cases}$$

- 汉明窗：

$$\omega(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0\le n\le N-1\\ 0, & otherwise\end{cases}$$

- 哈宁窗：

$$\omega(n)=\begin{cases}0.5-0.5\cos\left(\dfrac{2\pi n}{N-1}\right), & 0\le n\le N-1\\ 0, & otherwise\end{cases}$$

5.4.2 短时音频特征

短时物理特征主要包括：

- 短时平均能量
- 短时平均过零率
- 基音频率
- 子带能量率和频率质心
- MEL频率倒谱系数

1. 短时平均能量

信号 $\{x(m)\}$ 的短时平均能量定义为：

$$E_n = \frac{1}{N}\sum_{m=-\infty}^{\infty}[x(m)\times\omega(n-m)]^2$$

其中，N为帧长。

短时平均能量序列反映了声音信号振幅或能量随着时间缓慢变化的规律，对于语音信号来说，平时平均能量有一个重要作用是区分语音信号中的浊音成分和清音成分。

2. 短时平均过零率

短时平均过零率是语音信号时域分析中最简单的一种特征，其表达式如下：

$$Z_n = \frac{1}{2N}\sum_{m=-\infty}^{\infty}\left|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\right| \times \omega(n-m)$$

其中，N为帧长，sgn[ ]表示符号函数，即：

$$\mathrm{sgn}[x] = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

过零率有两类重要应用：第一，可用来粗略地描述信号的频率特征，即可粗略地估计频谱特性。第二，用于判别清音和浊音、静音与非静音。另外，语音和音乐的短时过零率曲线有较大的差别。

3. 基因频率

声音信号有和谐与不和谐之分，和谐声音可近似看作由一系列频率成整数倍关系的正弦波组成，其中最低的频率成分称为基音频率，相应的信号周期称为基音周期，其它频率成分都是基音频率的整数倍，称为谐波。

因此，通过检测声音信号中是否有连续稳定的基音频率(或基音周期)存在，可以区分声音信号是否和谐。

4. 子带能量率和频率质心

子带能量率和频率质心是两个频域特征，用于描述音频信号的频率分布。子带能量即音频信号在某一频带范围的能量，其定义为：

$$P_j = \int_{L_j}^{H_j} \left| F(\omega) \right|^2 d\omega$$

其中，子带j的频率范围为$\left[L_j, H_j \right]$。
子带能量率为音预信号在某一频带范围的能量占总能量的比率，其计算式为：

$$R_j = \frac{P_j}{P}$$

4. 子带能量率和频率质心

频率质心反映了音频信号频率分布的中心，其定义为：

$$\omega_c = \frac{\int_0^{\omega_0} |F(\omega)|^2 \, \omega d\omega}{\int_0^{\omega_0} |F(\omega)|^2 \, d\omega}$$

子带能量率和频率质心是两个相关的概念，它们都反映了音频信号的频率分布情况，而不同的音频信号类别的频率分布是不同的。

5. MEL频率倒谱系数

MEL频率倒谱系数（MFCC）被广泛地应用于语音识别和说话人识别中，它利用三角滤波器组对傅立叶变换能量系数滤波而得，且对其频域进行Mel变换，以更符合人类听觉特性。

MFCC系数的计算过程如下：

① 将信号分帧，预加重并进行加窗处理，然后进行短时傅立叶变换得到其频谱；

② 求出能量谱，并用M个Mel带通滤波器进行滤波，由于每个频带中分量的作用在人耳中是叠加的，因此将每个滤波器频带内的能量进行叠加，这是第k个滤波器输出功率谱为；

③ 计算离散余弦逆变换的MFCC系数。

$$C(n)=\sum_{k=1}^{M}\log x'(k)\cos[\pi(k-0.5)\,n/M] \quad n=1,2,\cdots,L$$

5.4.3 基于HMM的不良音频识别

1. 隐马尔科夫模型HMM

隐马尔科夫模型（HMM）是在马尔科夫链的基础上发展起来的，该模型是一个双重随机过程，不知道具体的状态序列，只知道状态转移概率，即模型的状态转换过程是隐蔽的，而可观察事件的随机过程是隐蔽状态转换过程的随机函数。这样站在观察者的角度，只能看到观察值，不像Markov链模型中的观察值和状态一一对应，不能直接看到状态，而是通过一个随机过程去感知状态的存在及其特性。因此称之为“隐”Markov模型，即HMM。

1. 隐马尔科夫模型HMM

一个HMM可以由下列参数描述：

① N：模型中的马尔科夫链状态数目。记N个状态为 $\theta_1,\cdots,\theta_N$ ，记t时刻Markov链所处状态为 q_t ，显然 $q_t \in(\theta_1,\cdots,\theta_N)$ 。

② M：每个状态对应的可能的观察值数目 $V_1,\cdots,V_M$ 。记M个观察值为 O_t，记t时刻观察到的观察值为 ，其中 $O_t \in(V_1,\cdots,V_M)$ 。

③ π：初始状态概率矢量，$\pi=(\pi_1,\cdots,\pi_N)$ ，其中

$$\pi = P(q_i = \theta_i),\quad 1 \le i \le N$$

④ A：状态转移概率矩阵，$A=(a_{ij})_{N\times N}$ ，其中

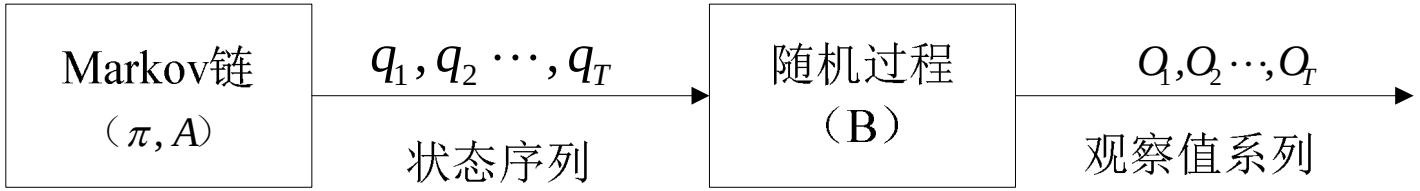
$$a_{ij} = P\left(q_{i+1} = \theta \middle| q_i = \theta_i\right),\quad 1 \le i, j \le N$$

1. 隐马尔科夫模型HMM

⑤ B：观察值概率矩阵，$B=(b_{ij})_{N\times M}$ ，其中

$$b_{jk}=P\left(O_i=V_k \,\middle|\, q_i=\theta_j\right),\quad 1\le j\le N, 1\le k\le M$$

这样，一个HMM就可以记为：$\lambda=(N,M,\pi,A,B)$ ，或简记为 $\lambda=(\pi,A,B)$ 。HMM可以形象地描述为下图所示的两部分。



2. 基于HMM的不良音频内容识别

不良音频通常表现为两种形式：不良的女音、语音对话中的内容为不良等。不良的女音识别过程为：先对可疑音频段进行提取，考察不良女音与其他音频段在短时特征及段特征上的区别，选择参量构造段特征向量。而对于语音对话中的不良内容则需要进行识别语音，得到语音文字内容后采用不良文本内容识别方法进行识别。

2. 基于HMM的不良音频内容识别

（1）可疑音频段的提取

通过分析可疑音频段发现，可疑音频段的持续时间一般在200ms-4000ms之间，以静音段相隔，且具有反复性和持续性。对于女音来说，它主要反映浊音部分的特征；过零率定义为一帧信号中波形穿越零电平的次数，对于语音，它主要反映清音部分的特征。

2. 基于HMM的不良音频内容识别

主要步骤为：

①滤波：人类语音主要在300-3400Hz，所以在检测之前，首先将音频信号通过带通滤波器过滤不关心的音频部分。

②归一化：将音频信号归一化后，即可为短时能量和过零率分别确定高低两个门限用于音频段端点检测。

③端点检测：包括起始点与结束点的检测和判断。加入低门限判断的目的是减少微弱声音，比如背景噪声对分段的影响，通过两个门限的控制，可以得到较好的分段结果。

④保留可疑音频段：成功标记出起始、结束点之后，将段持续时间与所设定的最短、最长时间段门限Tthr1、Tthr2比较，以排除噪声、长段对话、音乐等不相关内容，而只保留所关心的可疑音频段。

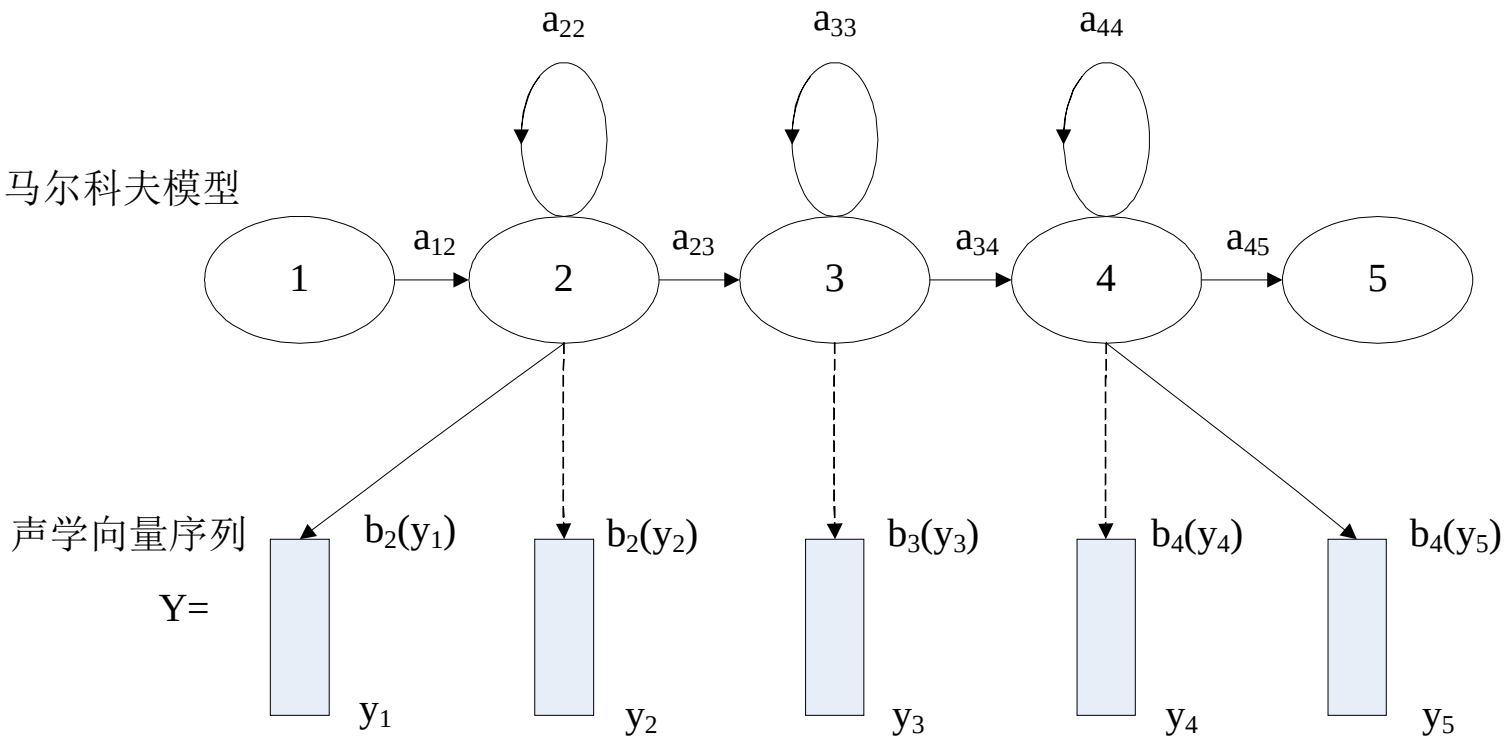
2. 基于HMM的不良音频内容识别

（2）语音识别

若不良音频是语音对话中的内容为不良，则需要进行语音识别。对于一个语音特征序列W，通常将其分解成为一个音素（Phoneme）所组成的序列。为了适应不同可能的发音带来的变化，似然　　可以由多个不同的发音来计算得到：

$$p(Y|W)=\sum_{Q} p(Y|Q)p(Q|W)$$

2. 基于HMM的不良音频内容识别



2. 基于HMM的不良音频内容识别

如图所示，每一个音素q可以用一个连续密度的隐马尔科夫模型来表示，模型包含状态转移矩阵$\{a_{ij}\}$和输出观察概率分布$\{b_j()\}$。通常输出观察概率分布是用高斯混合模型来描述：

$$b_j(y)=\sum_{m=1}^{M} c_{jm}\mathrm{N}(y;u_{jm},\Sigma_{jm})$$

其中N表示一个均值为u_{jm}，协方差为Σ_{jm}的正态分布。因为特征向量y的维数相对较高，协方差矩阵一般都约束为对角阵。给定以音素为基本单位的复合的HMMQ，对应声学模型的似然函数为：$p(Y|Q)=\sum p(X,Y|Q)$

其中$X=[x(0),\cdots,x(T)]$是一个基于复合模型的状态序列，并且

$$p(X,Y|Q)=a_{x(0),x(1)}\prod_{t=1}^{T} b_{x(t)}(y_t)a_{x(t),x(t+1)}$$

2. 基于HMM的不良音频内容识别

将HMM模型应用于语音识别系统中，需要解决三个基本问题：

①已知观察序列y和模型 $\lambda=(\pi,A,B)$ ，计算由此模型产生观察序列的概率 $p(y\mid\lambda)$ ，一般采用“向前-向后”算法。

②已知观察序列y和模型λ，确定一个合理的状态序列，使之能最佳的产生y，即如何选择最佳的状态序列Q，通常采用Viterbi算法。

③根据观察序列不断修正模型参数(π,A,B)，使 $p(y\mid\lambda)$ 最大，通常采用Baum-Welch算法。