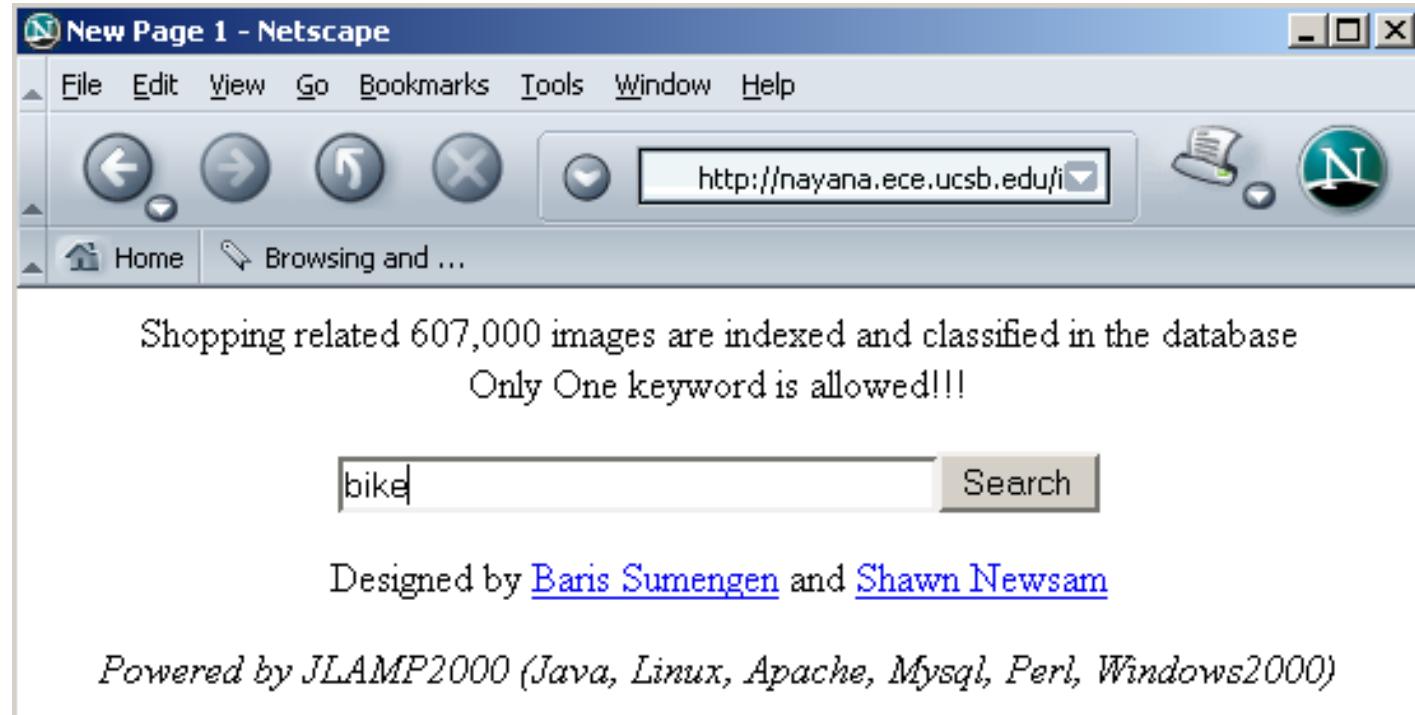


# Web信息处理与应用

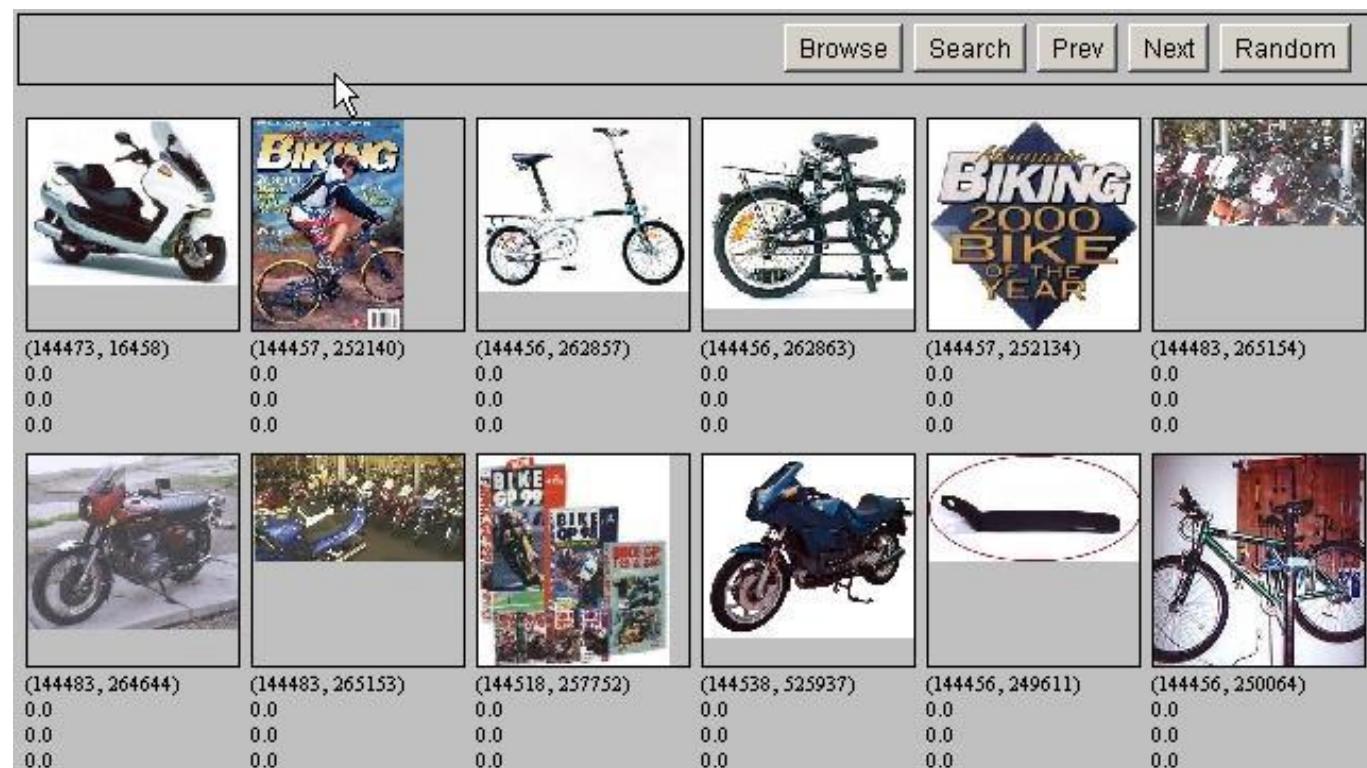
## 第八节 多模态检索（综述）

徐童 2021.11.1

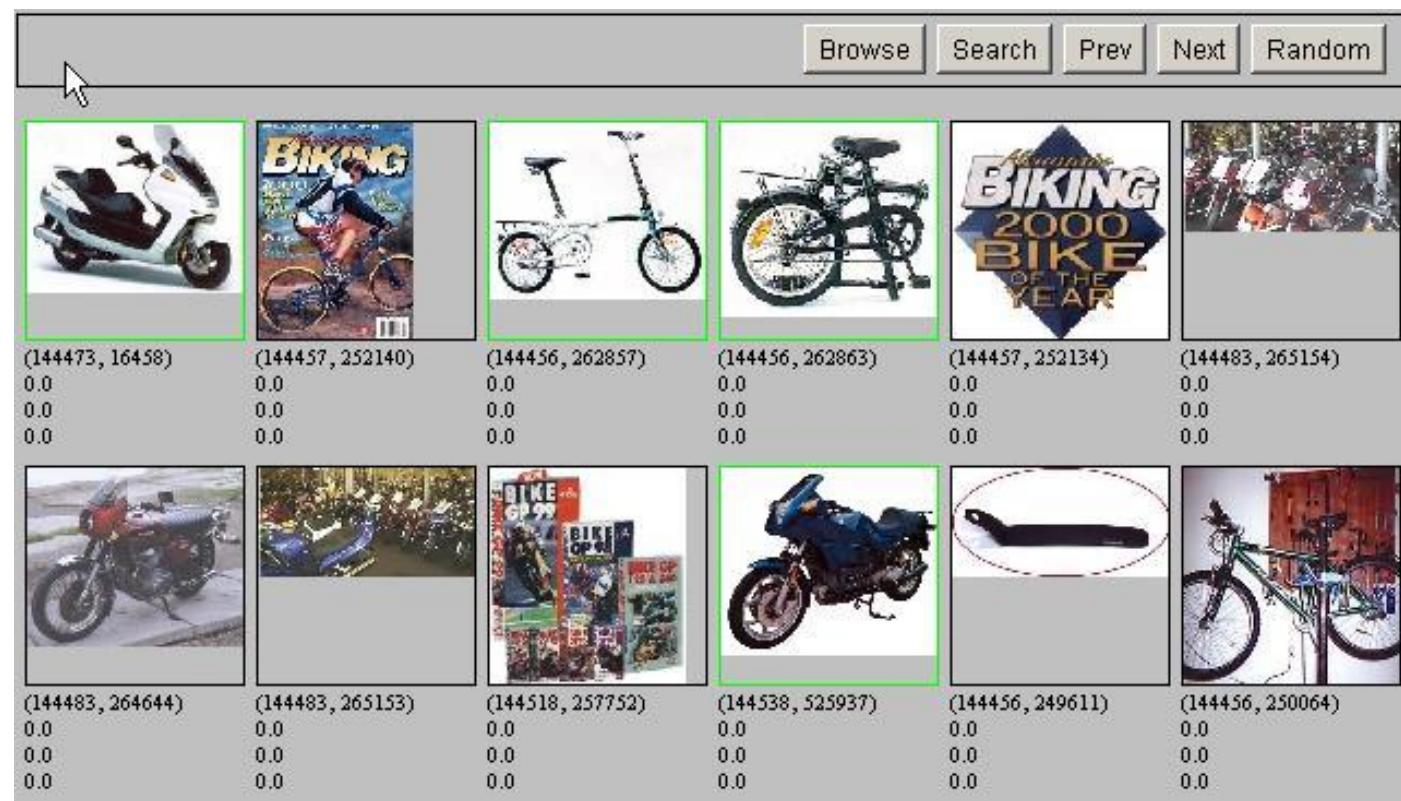
- 其实，我们早就见识过多模态检索
- 用户的初始查询需求：搜索“Bike”相关的图片



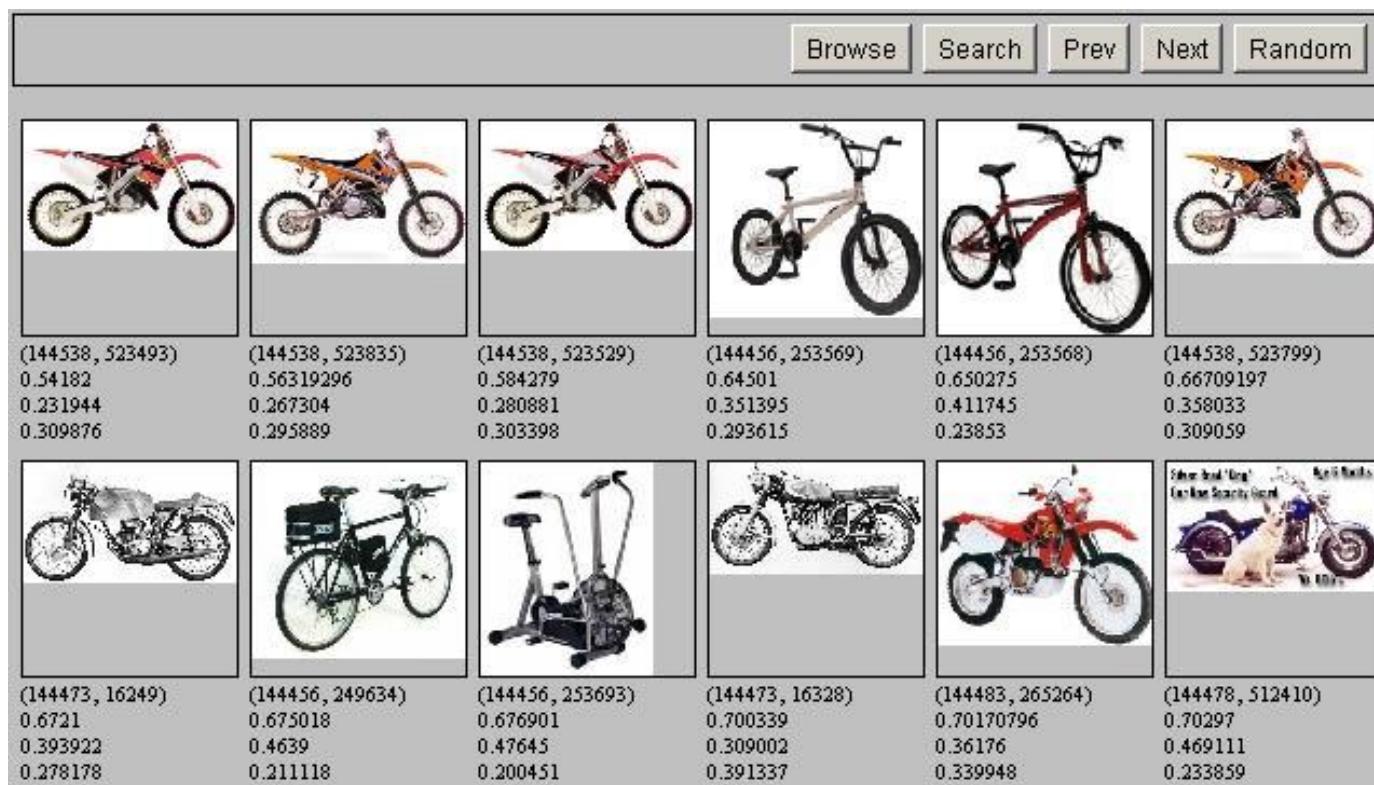
- 基于查询条件的初始检索结果



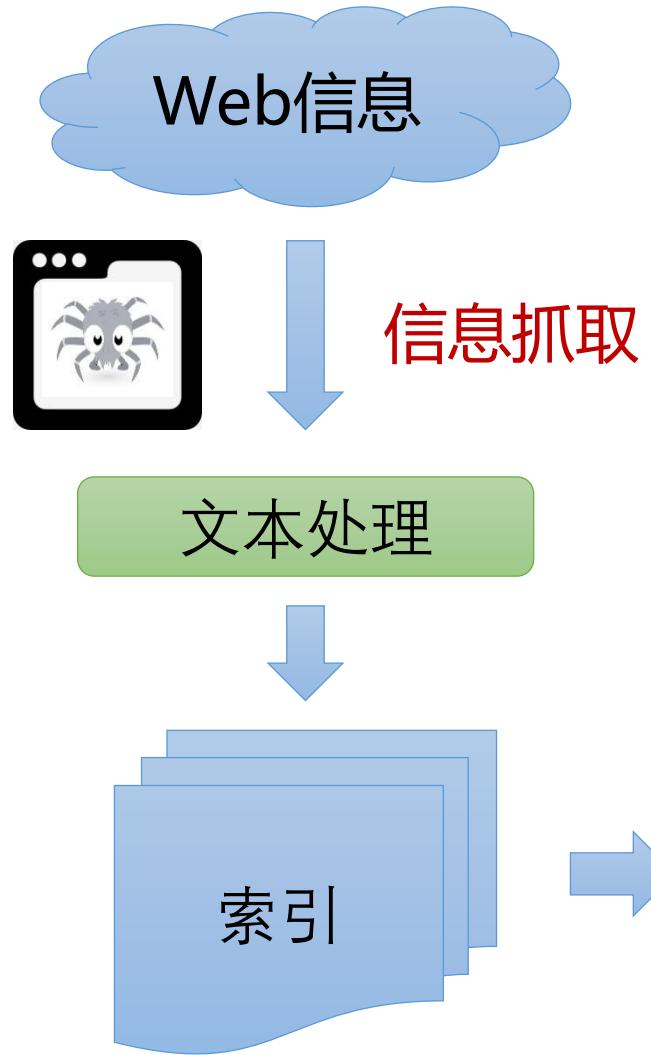
- 用户对部分相关图片进行了标记（或点击等行为）



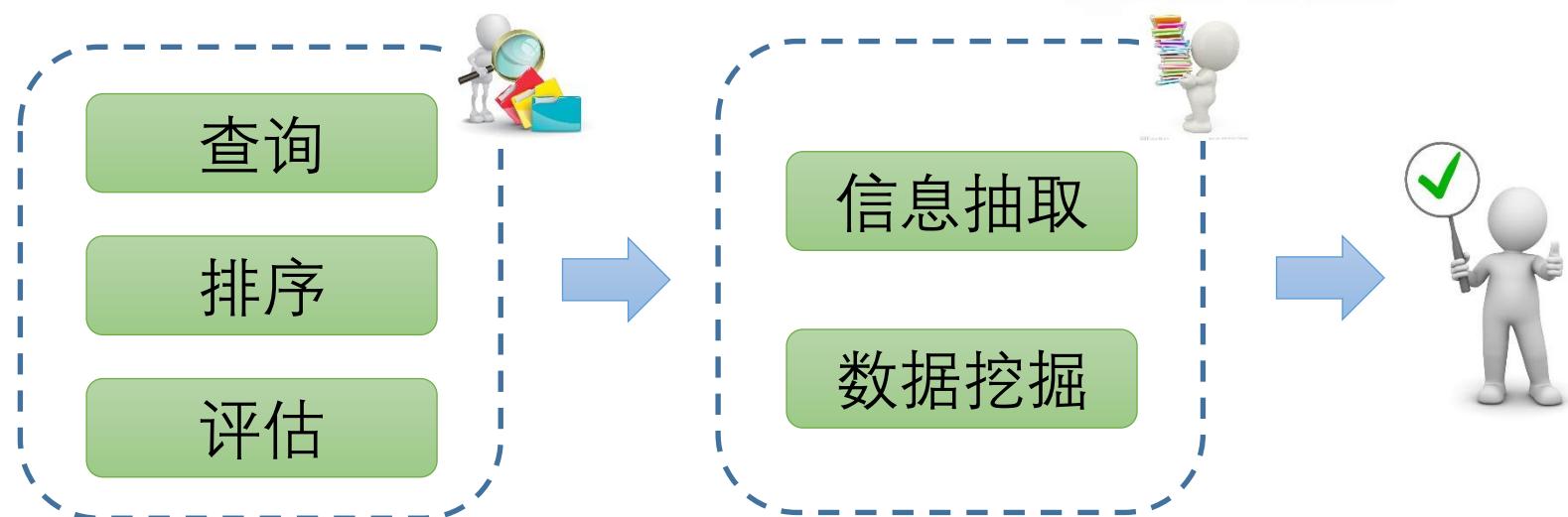
- 基于用户反馈，得到了更新的搜索结果，以车型图片为主



- 本课程所要解决的问题



## 第七个问题： 多模态信息，如何检索？



- **什么是多模态搜索?**
- 面向 “多媒体文档” 的搜索技术 / 系统
  - 多媒体文档：包含多种模态的信息，如文本、图像、视频、音频等
  - 从广义上说，只要是面向非文本信息的搜索都可以算多模态搜索



- **什么是多模态搜索?**

- 由于“广义概念”的存在，多模态搜索系统可以简单分为以下几类
  - 面向单一模态的检索
    - 图片搜索（以图搜图）、音频检索（曲调识别）等
  - 跨模态检索，如借助文字标签搜索图片
  - 真·多模态搜索
    - 建立在多模态特征融合基础之上的搜索任务

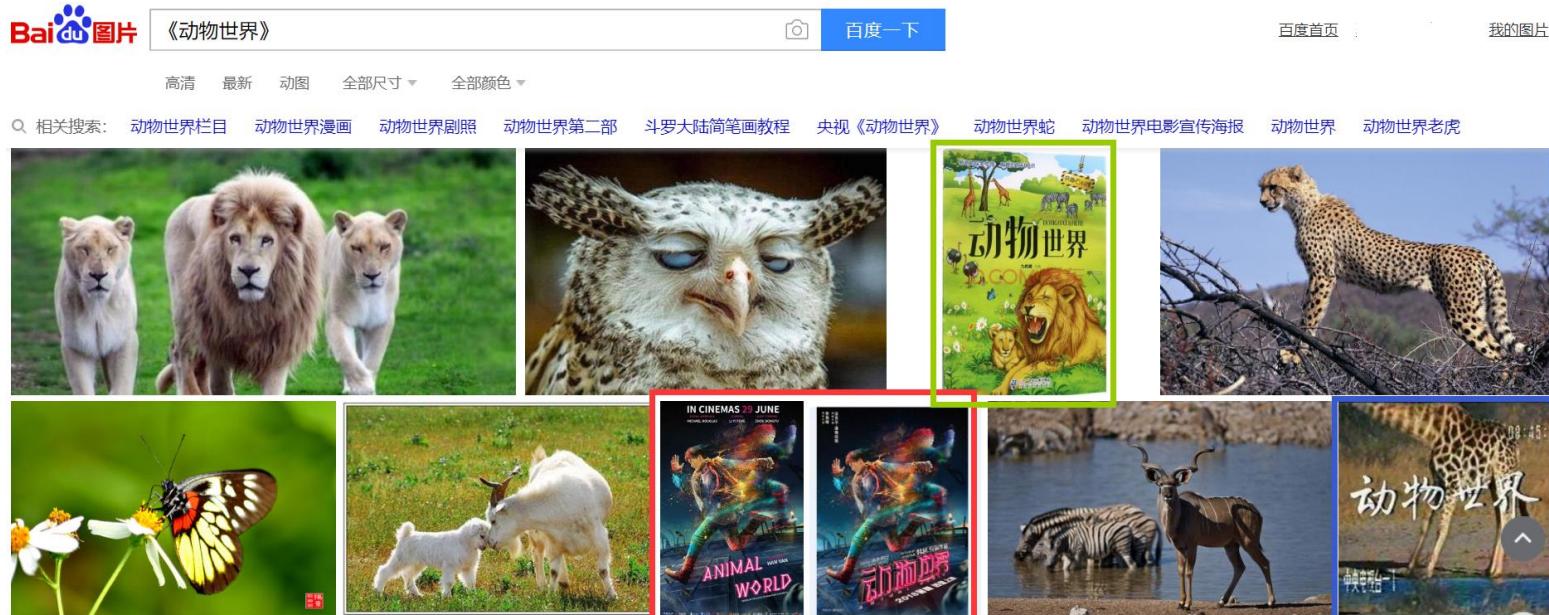


- 什么是多模态搜索?

- 常见的多模态搜索

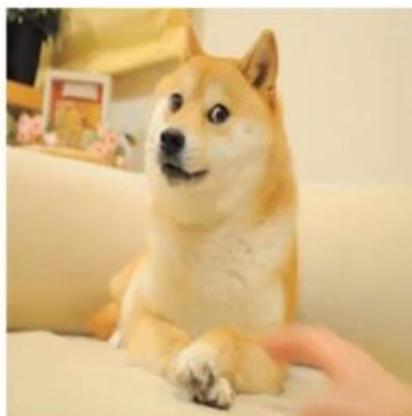


- 多模态信息的特点
- 虽然多媒体文档可能包含关键词，但难以涵盖其中非文字信息的真正内涵



不同含义的“动物世界”，你要找的是哪一种？

- 多模态信息的特点
- 潜藏在文本语义与多模态信息之间的“语义鸿沟”（Semantic Gap）问题



			Blue channel										
			Green channel										
			Red channel			24	56	230	1	...	8	39	
1	120	67	89	107	...	13	18	8	...	12	18	81	71
2	12	216	145	26	...	181	81	71	...	...	56	...	7
3	0	16	4	45	...	44	56	...	...	...	12	...	12
4	0	78	90	167	...	25	...	...	...	...	...	...	7
...	...	...	...	...	...	...	...	...	...	...	...	...	...
64	12	67	82	141	...	12	12	12	12	12	12	12	12
	1	2	3	4	...	64							

Image array: [64 x 64 x 3]

计算机“看”到的：  
An image is just a big grid of numbers  
between [0, 255]  
e.g. 64 x 64 x 3 (3 channels RGB)



Human Perception

- **多模态检索的基本方法**
- 总体而言，多模态检索可笼统地划分为特征/语义两种方向的检索
  - **基于特征的检索：**对多媒体文件本身进行特征刻画
    - 图像特征：颜色、纹理、形状……
    - 音频特征：音高、音调、频率……
  - **基于语义的检索：**通过多媒体文件对应的文本/标签信息进行检索
    - 元数据、标签/概念、事件、行为、空间关系……



- 面向图像的检索

- 基于图像内容的检索

- 基于文本信息的检索

- 面向视频的检索

- 面向音频的检索

- 多模态混合检索

## • 基于内容的检索

- Content-based Image Retrieval (CBIR)
  - 允许用户输入一张图片，以查找具有相同或相似内容的其他图片。
  - 输入的图片往往被称作样例图 (Query-by-Example, QBE)
  - CBIR中的内容，往往指图像或视频的特征描述
    - 基于这一特征描述，采用数学模型衡量相似性
    - 用户可根据满意度进行反馈和结果修正

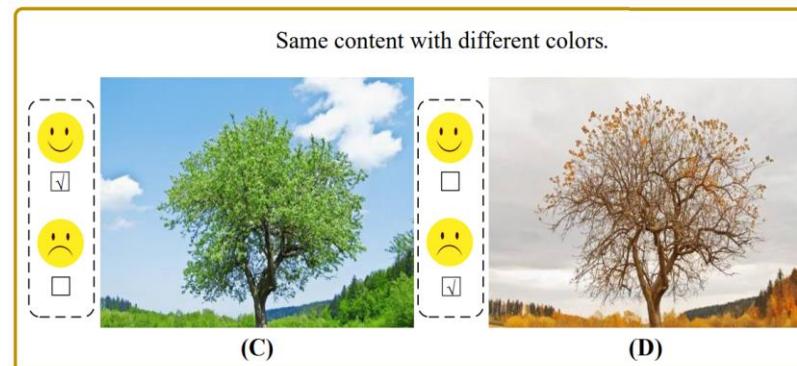
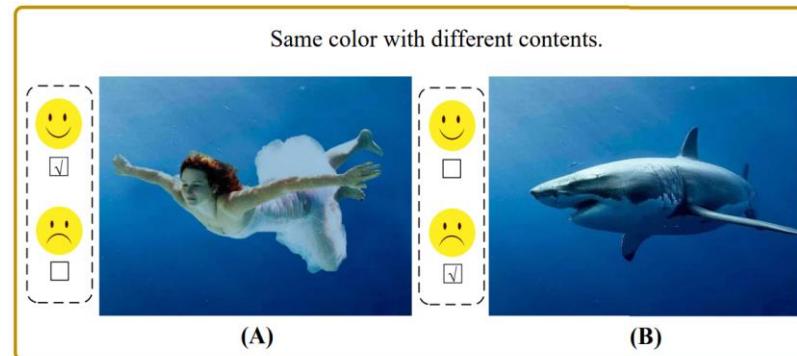


## • 图像检索的基本类型

- 具体到图像查询而言，主要依据图像的颜色、纹理、形状等进行查询
  - 颜色查询：查询与用户所选图片颜色（或颜色分布）相似的图像
    - 颜色特征是最为可靠和通常的视觉特征
  - 形状查询：用户给出某一形状或勾勒草图，利用形状特征进行检索
  - 纹理查询：用户给出包含某种纹理的图像，查询含有相似纹理的图像
    - 纹理包含了关于表面的结构布局及周围环境等信息

- **图像检索的基本类型**

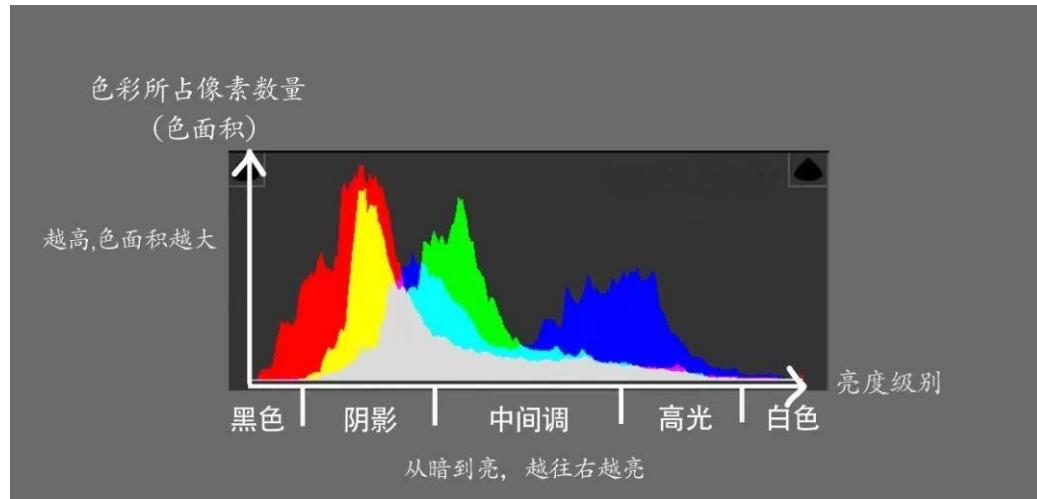
- 具体到图像查询而言，主要依据图像的颜色、纹理、形状等进行查询
  - 小拓展：当然，还有许多要素可以用于描述图片，e.g., 情感因素



←相似的配色/图案可能表达不同的情感

## • 基本手段（1）：颜色特征

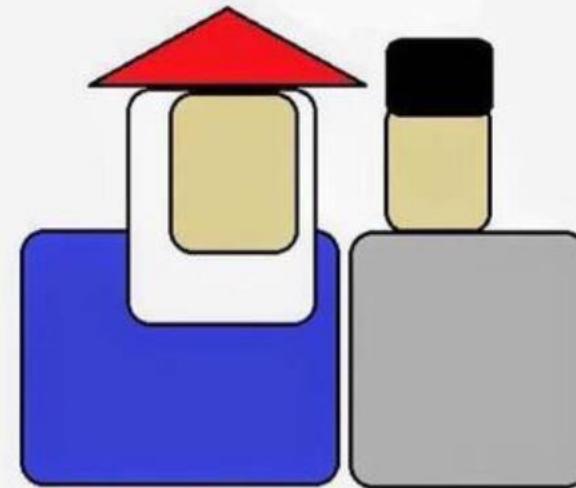
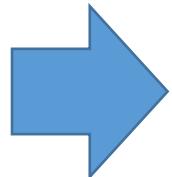
- 颜色直方图（Color Histogram）
- 横轴表示颜色或颜色等级，纵轴表示在某个颜色等级上，具有这种颜色的像素在整个图像中所占的比例。



- **基本手段（1）：颜色特征**

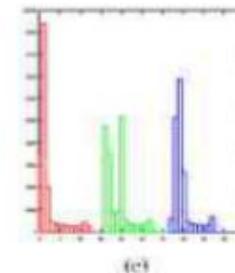
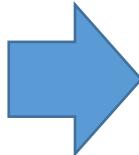
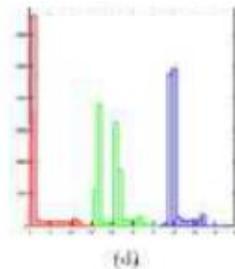
- 基于颜色直方图的检索

- 颜色组成查询：例如查询“大约30%红色，50%蓝色的图像”。
- 示例查询：基于用户输入示例，计算颜色直方图并进行查询。



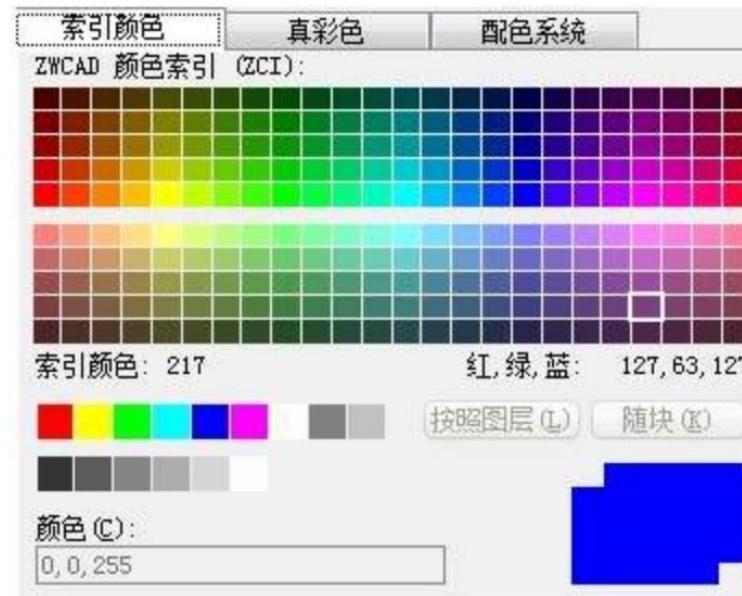
## • 基本手段（1）：颜色特征

- 颜色直方图的优缺点
- 只关心颜色的比重，而不关心颜色的位置，因此无法描述图像中的物体。
  - 更适合用来描述那些难以自动分割的物体。
  - 另外，当颜色数量过多的时候，必须减少颜色的级数。



## • 基本手段（1）：颜色特征

- 颜色矩（Color Histogram）
- 当颜色种类过多时，颜色直方图会很冗长，而颜色矩可以减少颜色特征。
  - 早期256色 → 如今的16/24/32位（其实是24位+8位灰度）



- **基本手段（1）：颜色特征**

- 颜色矩 (Color Histogram)
- 当颜色种类过多时，颜色直方图会很冗长，而颜色矩可以减少颜色特征。
  - 每种颜色采用三阶特征（均值、方差、斜度），一般考虑三种原色。

$$\mu_i = \frac{1}{N} \sum_{j=1}^N p_{i,j}$$

$$\sigma_i = \left( \frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}}$$

$$s_i = \left( \frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^3 \right)^{\frac{1}{3}}$$

## • 基本手段（1）：颜色特征

- 颜色直方图与颜色矩都具有的缺陷：难以描述颜色空间分布
- 解决方法：引入空间位置信息
  - 启发式方法1：将直方图一分为二，把颜色连贯的像素与不连贯的像素进行分开统计
  - 启发式方法2：将图像划分为子块，根据子块间颜色特征的相似度的加权和来计算相似度
    - 如基于N叉树的颜色布局方法，或对图像进行分割

- **基本手段（2）：纹理特征**
- 纹理（Texture）是物体表面在视觉上的一种先天特征
  - 不同的材质，如纸张、木头、纤维等都有不同的纹理
  - 一幅图像可以视作由多个不同纹理区域拼接而成
    - 每个区域的纹理相当于这个区域的代表
    - 基于纹理可实现辨识图像的目的



- **基本手段（2）：纹理特征**

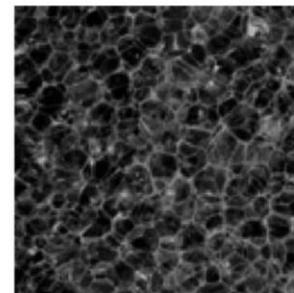
- Tamura纹理特征：用于描述人对纹理的视觉感知特性

Tamura et al., Textural Features Corresponding to Visual Perception, 1978

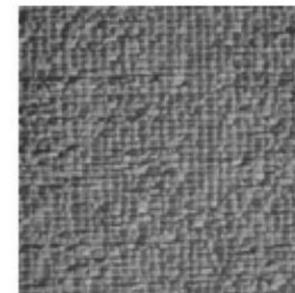
- 粗糙度：反应纹理粒度，主要衡量基元尺寸，颗粒越大则越粗糙
- 对比度：主要衡量灰度的变化范围、幅度、边缘锐度等信息
- 方向性：描述纹理如何沿着某些方向发散或者集中的
- 近年来，基于小波变换、傅里叶变化等新技术的方法得以广泛应用

## • 基本手段（2）：纹理特征

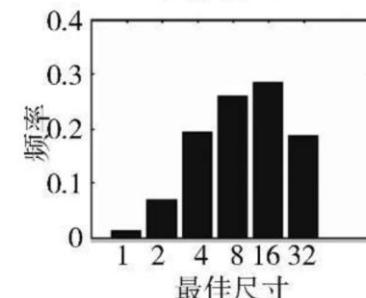
- Tamura纹理特征：用于描述人对纹理的视觉感知特性
  - 粗糙度案例：反应纹理粒度，主要衡量基元尺寸，颗粒越大则越粗糙



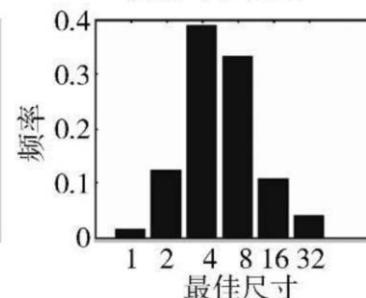
a 泡沫纹理



b 酒椰纤维纹理



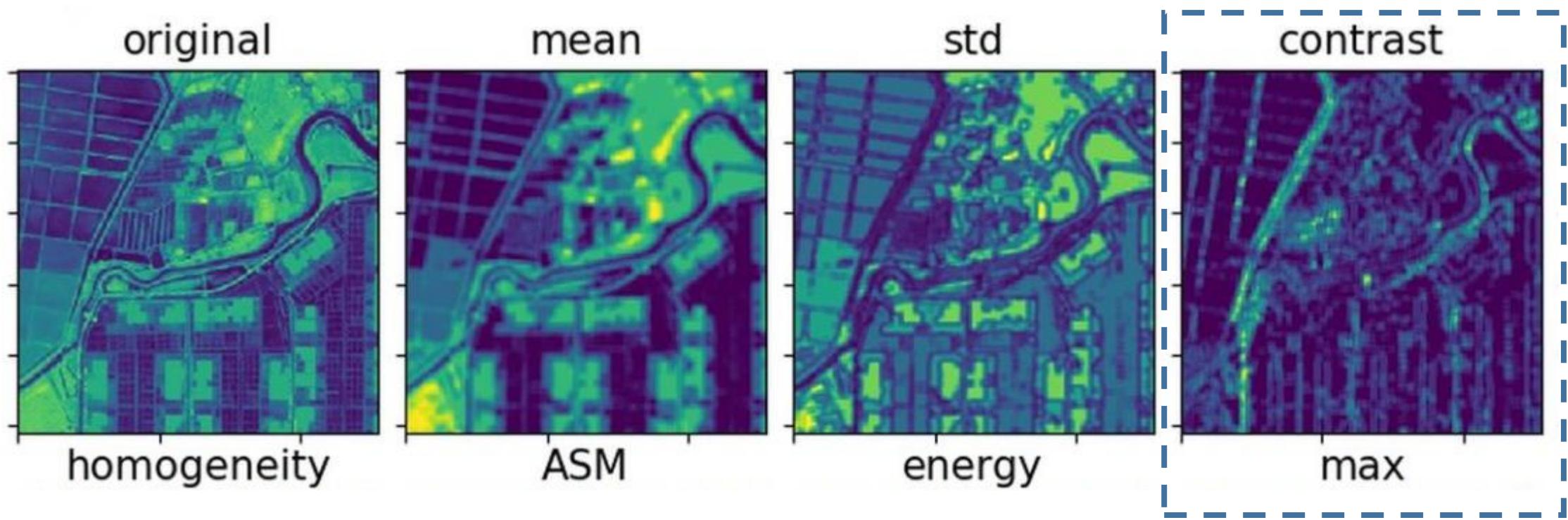
c 泡沫纹理粗糙度直方图



d 酒椰纤维纹理粗糙度直方图

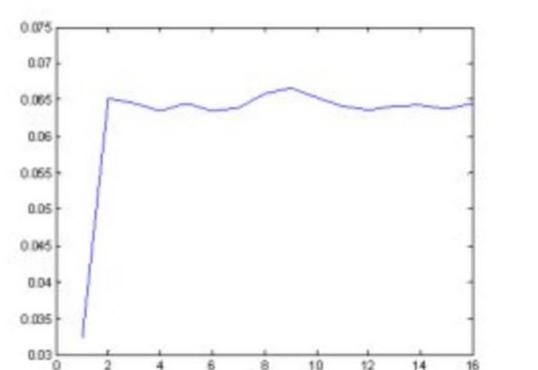
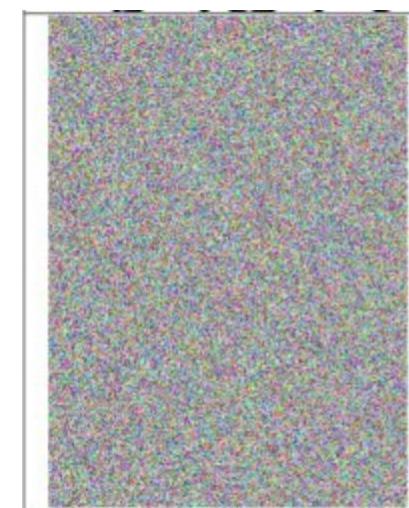
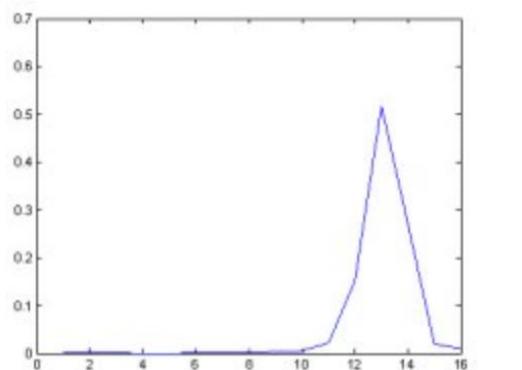
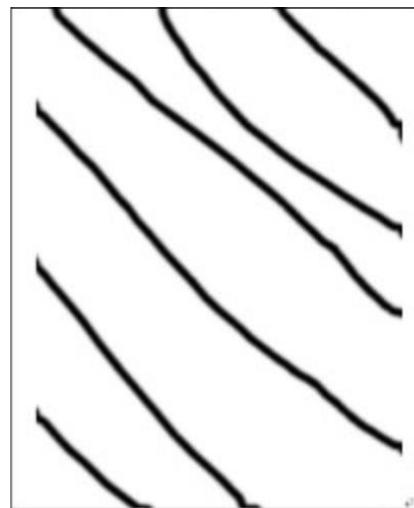
## • 基本手段（2）：纹理特征

- Tamura纹理特征：用于描述人对纹理的视觉感知特性
  - 对比度案例：主要衡量灰度的变化范围、幅度、边缘锐度等信息



## • 基本手段（2）：纹理特征

- Tamura纹理特征：用于描述人对纹理的视觉感知特性
  - 方向性案例：描述纹理如何沿着某些方向发散或者集中



- 基本手段（2）：纹理特征

- 纹理特征有着广泛的用途



## • 基本手段（3）：形状特征

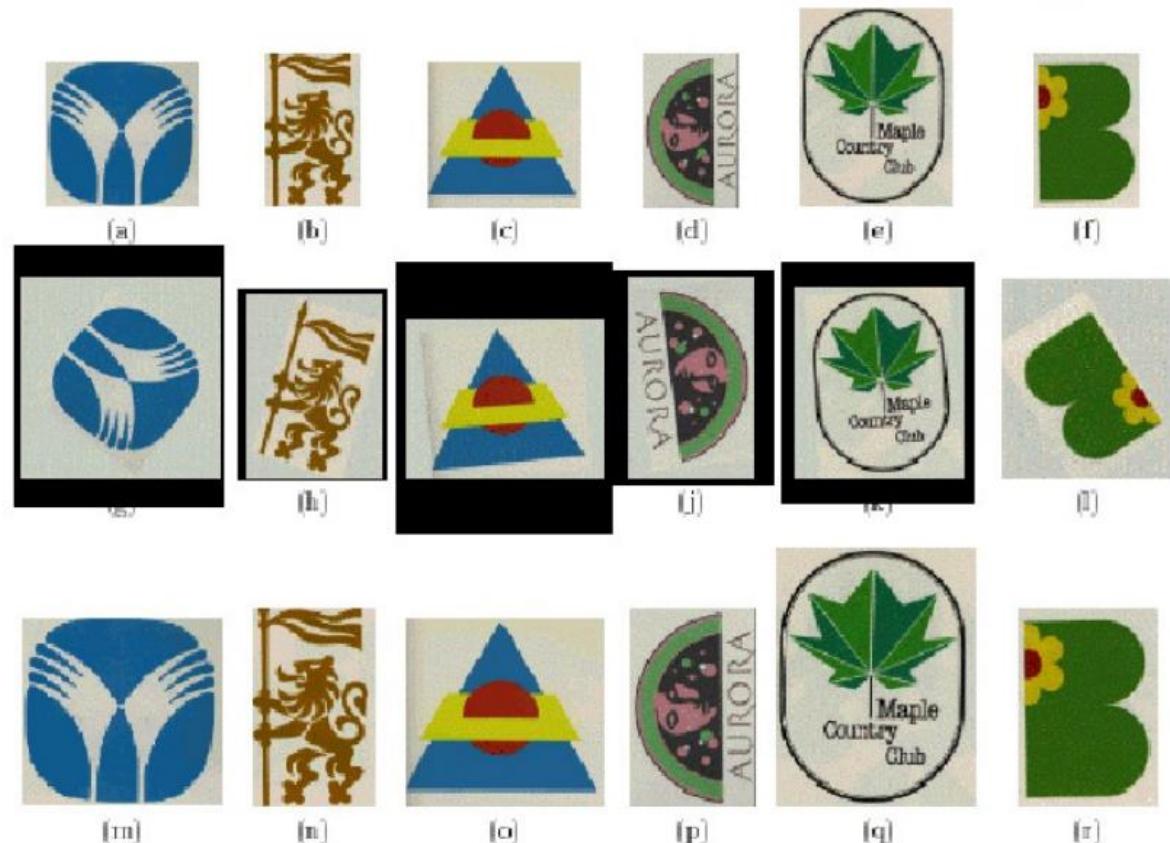
- 定义：图像中所包含对象或感兴趣区域的形状，用于对于图像内容更细粒度的描述
- 基础：对图像的分割和边缘提取
  - 边缘指图像中其周围像素灰度（颜色）有阶跃变化的像素集合
  - 提取方法包括滤波器、微分法等
- 难点：相似性度量比较困难
  - 形状的不规则性，难以用特定指标加以衡量



- 基本手段（3）：形状特征

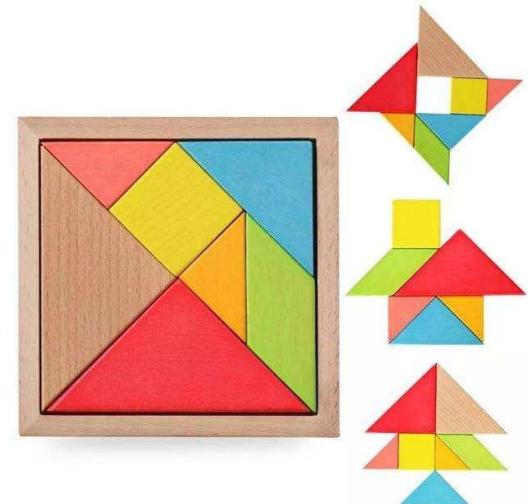
- 形状特征的变换不变性

- 长/短轴比
- 周长/面积比
- 最近与远点连线之间的夹角
- .....



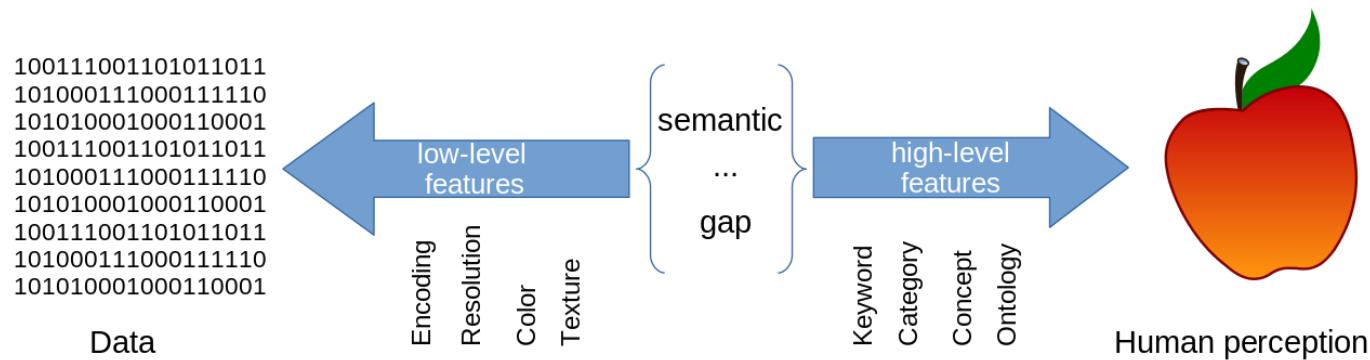
- **基于内容的图像检索：相似性度量**

- 对于图像的相似性衡量，单一特征往往难以有效衡量
  - 一般而言，采用多种特征的加权和进行整体相似性度量
- 同时，除了图像的整体相似性，还要考虑各个区域之间的相似性，甚至各个区域之间空间关系的相似性
  - 例如，由若干模块组成的图像，需考虑模块的空间关系



## • 基于内容的图像检索：相似性度量

- 整体而言，CBIR试图从信号处理角度入手，使检索过程符合人类的视觉特性（所见即所得）
- 然而，CBIR面临着前面所述的“语义鸿沟”（Semantic Gap）问题
  - 低层视觉特征与高层语义特征不存在直接联系
  - 特征相似 ≠ 语义相似



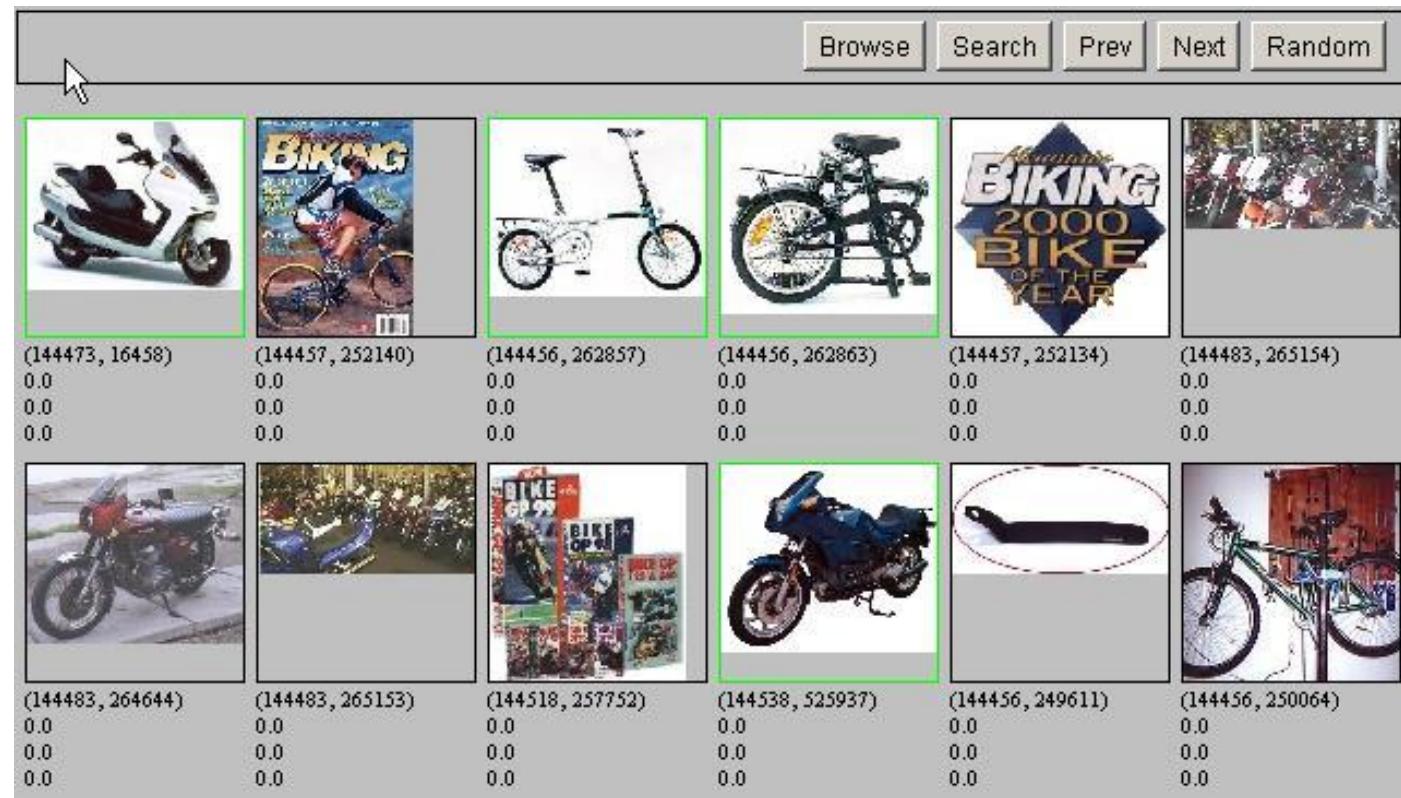
- 基于内容的图像检索：相似性度量

- 整体而言，CBIR试图从信号处理角度入手，使检索过程符合人类的视听觉特性（所见即所得）
- 然而，CBIR仍面临巨大的“语义鸿沟”（Semantic Gap）问题
  - 例如，机器可以识别包含狗的图片，却不能理解何为“狗”



- 一种挽救的办法：相关反馈与学习功能

- 通过人机交互，可以帮助系统了解用户对结果的满意程度和真实需求



没错又是这张图

- 面向图像的检索

- 基于图像内容的检索

- 基于文本信息的检索

- 面向视频的检索

- 面向音频的检索

- 多模态混合检索

- 基于文本的图像检索

- 更为直接、更为便利的查询方式，同时也是更为传统的查询方式
  - 依赖于对于图像/视频信息的高质量语义描述

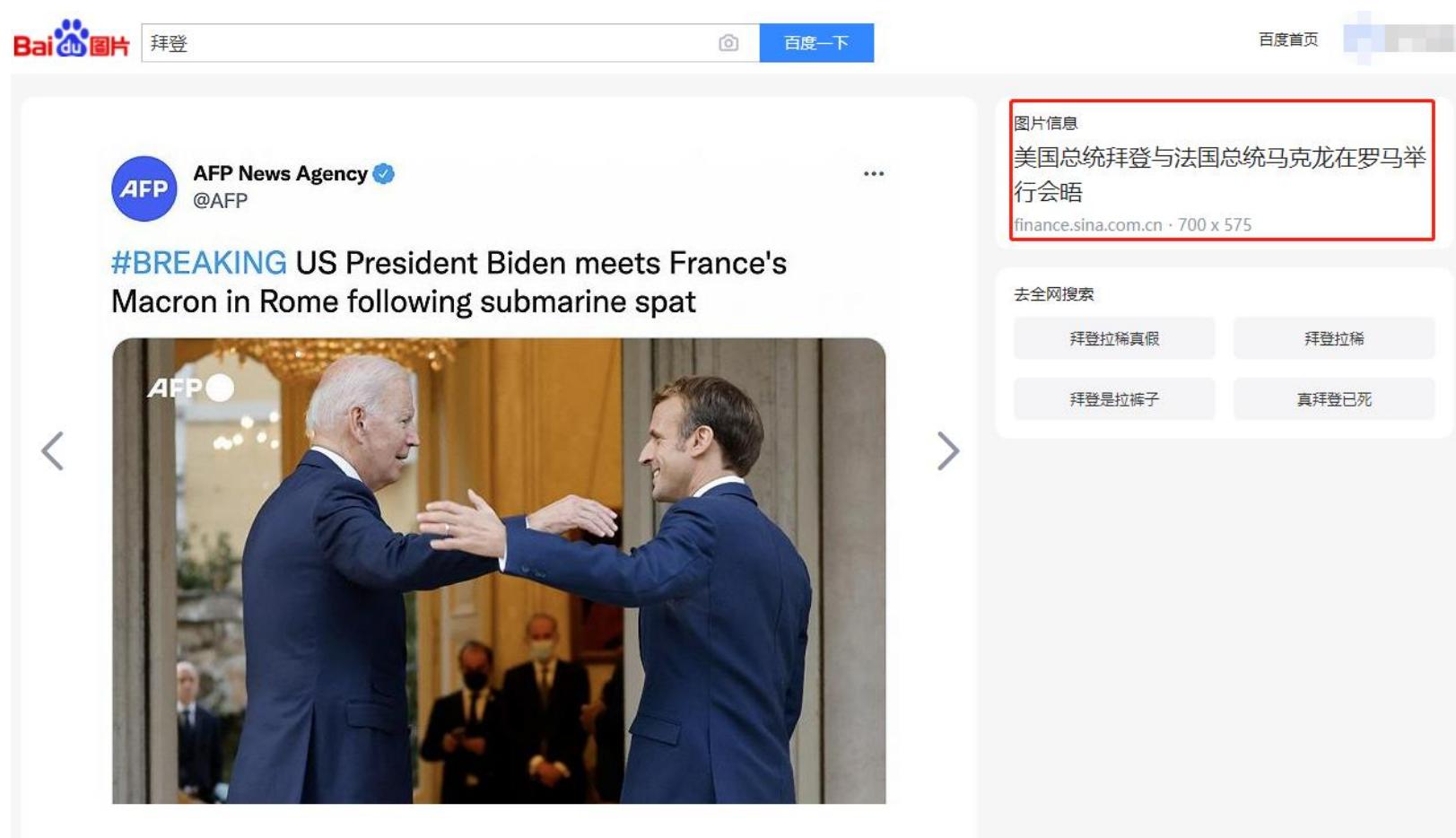


## • 文本信息来源 (1) : 手工标注

- 可以保证文本的代表性与相关性，但完全保证不了效率
  - 几乎是Mission Impossible，类似1994年的Yahoo!
  - 即使对于同一幅图像，不同的人也可能会有不同的理解与描述



- 文本信息来源（2）：元数据分析
- 抓取图片时，同时获取的其他信息，如链接文字、标题、关联页面等



## • 先前内容回顾：HTML中的重要部分

- <head><title> text </text></head>
- 是搜索服务显示的内容之一（URL，标题，摘要等）
- 
- 图片描述，可以帮助我们做“从文字到图片”的查询
- <a href="url" title="text">link text</a>
- 有助于理解目标网页内容及网页之间在内容上的关系



- 文本信息来源（2）：元数据分析

- 元数据可能缺失，即使有，也未必与图像内容相关

- 如下图，特朗普在哪儿？



图片信息

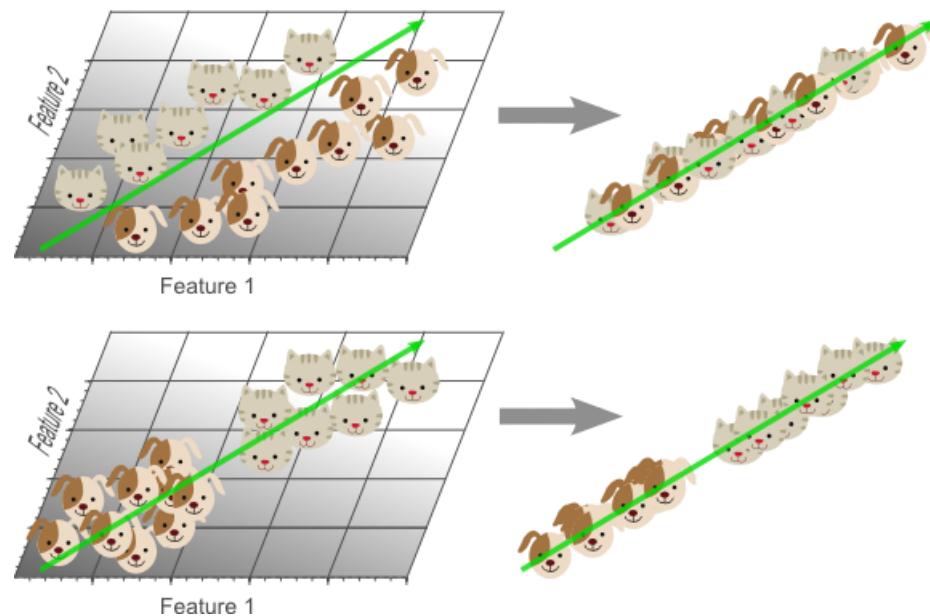
特朗普打算铤而走险?拒绝将白宫让给拜登,双方准备硬碰硬  
new.qq.com · 1000 x 667

去全网搜索

拜登全民英语 拜登大小便失禁  
为何拜登比特朗普更坏 拜登吃婴儿

- 文本信息来源（3）：算法标注

- 通过各种机器学习的算法，对图像进行有监督/无监督的标注
  - 本质上可视作一个（多）分类问题



## • 文本信息来源（3）：算法标注

- 图像的标签与识别，都可以转化为分类问题加以求解
  - 算法输入：图像（向量化）特征描述
    - 可采用图像搜索的各种特征，也可结合图像的Metadata进行描述
  - 算法输出：一个或多个分类标签
- 缺陷：往往只能表示简单、笼统的图像语义，缺乏更深层的信息
  - 例如：对象空间关系语义、动作与事件语义等

- 文本信息来源（3）：算法标注

- 空间关系与事件语义等深层语义信息，往往需要在准确识别图像中的实体的基础之上，进行一定的推理。
  - 例如，图像中爱因斯坦和居里夫人中间的人是谁？（答：洛伦兹）



## • 文本信息来源 (3) : 算法标注

- 空间关系与事件语义等深层语义信息，往往需要在准确识别图像中的实体的基础之上，进行一定的推理。
  - 事实上，类似的推理问题不仅在检索，而且在其它领域也有广泛应用



Is the **bowl** to the right of the **green apple**?

What type of **fruit** in the image is **round**?

What color is the **fruit** on the right side, **red** or **green**?

Is there any **milk** in the **bowl** to the left of the **apple**?

## • 从图像标注到图像描述

- Image Caption, 基于给定的图片，自动生成一段描述性文字。
  - 类似于小学作文练笔时的“看图说话”
  - 难点在于：不仅要检测物体，还要理解物体间的关系



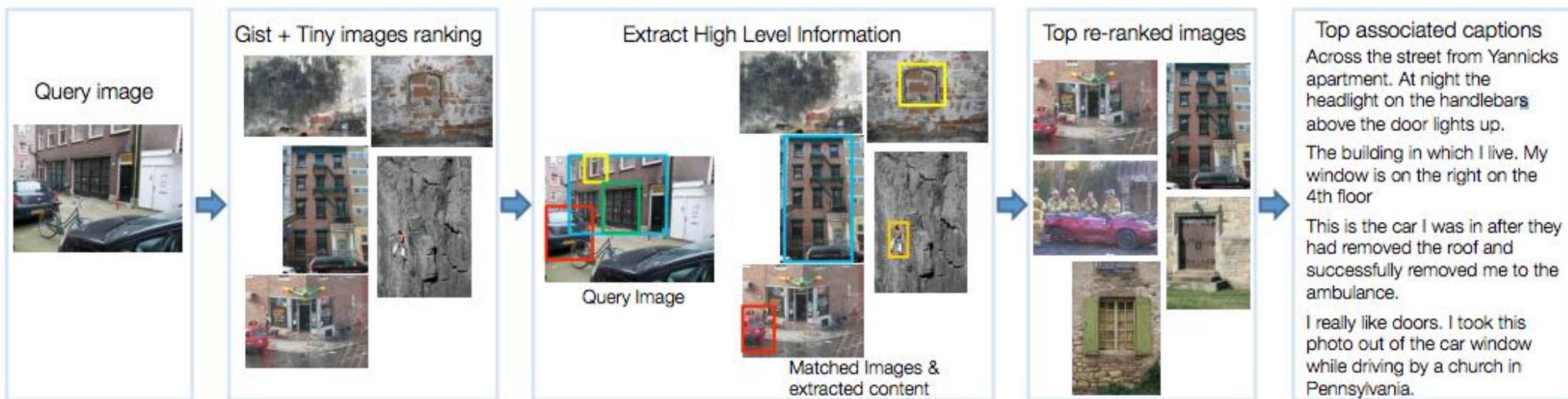
The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.

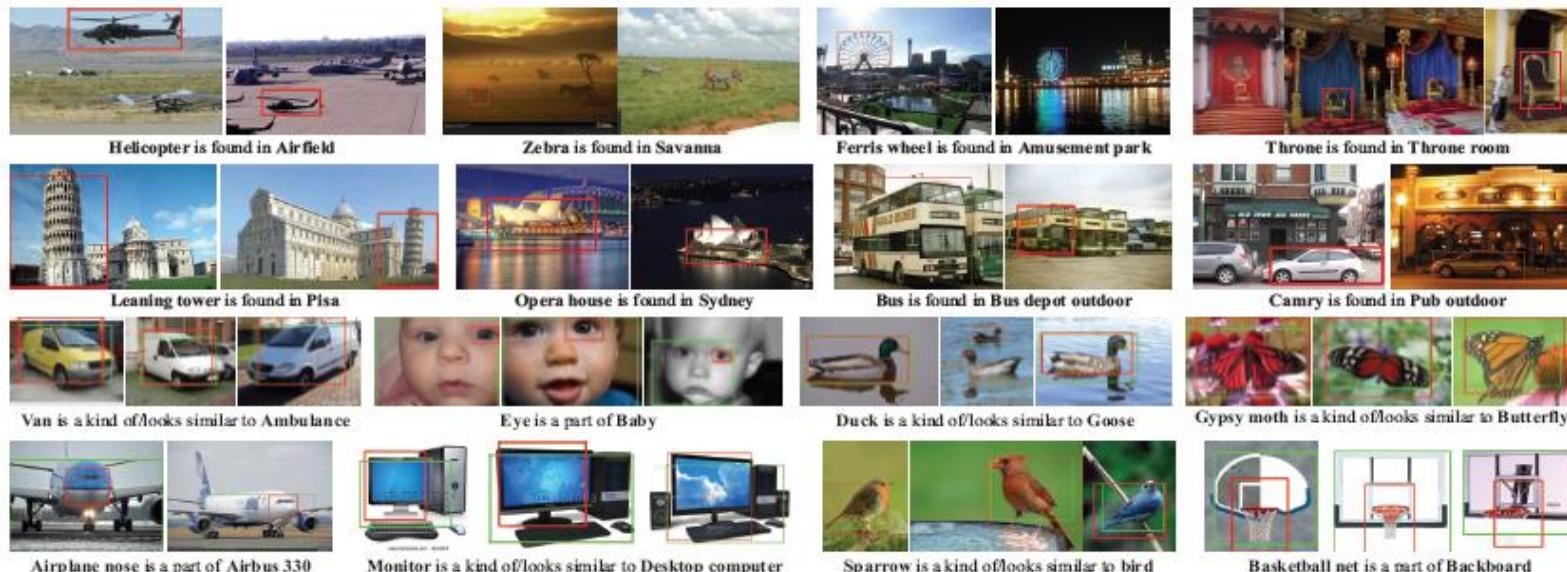
## • 从图像标注到图像描述

- 早在2011年，学者们即开始研究如何基于已有的图片即描述数据集，来为新图片生成相应的描述
  - 在图片全局匹配与高维图片信息抽取的基础之上，对图像重排序并生成描述



## • 从图像标注到图像描述

- 2013年，CMU提出了名为NEIL的知识抽取程序，持续标注图片数据，并从中自动抽取实体间关联
  - NEIL = Never Ending Image Learner



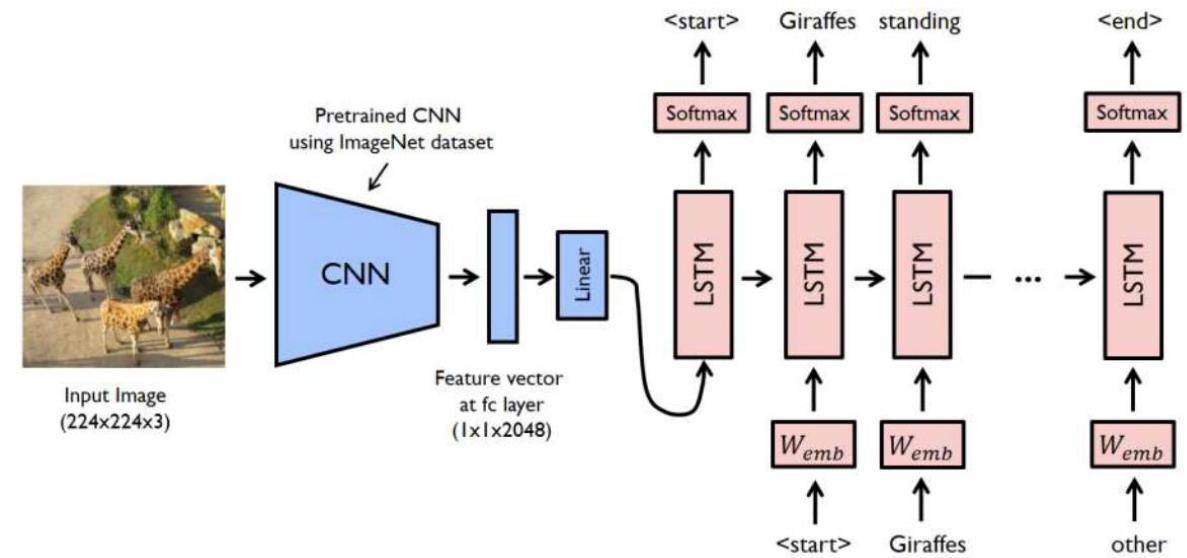
NEIL: Extracting Visual Knowledge from Web Data – Chen X et al, ICCV 2013.

## • 从图像标注到图像描述

- 2015年，图像描述的经典之作Show and Tell诞生
  - 借鉴了机器翻译的思想，但将NMT中处理句子的RNN换成了处理图像的CNN
  - 在此基础之上，又衍生出了基于注意力机制提升的Show, Attend and Tell



A group of people shopping at an outdoor market.  
There are many vegetables at the fruit stand.



# 图像检索

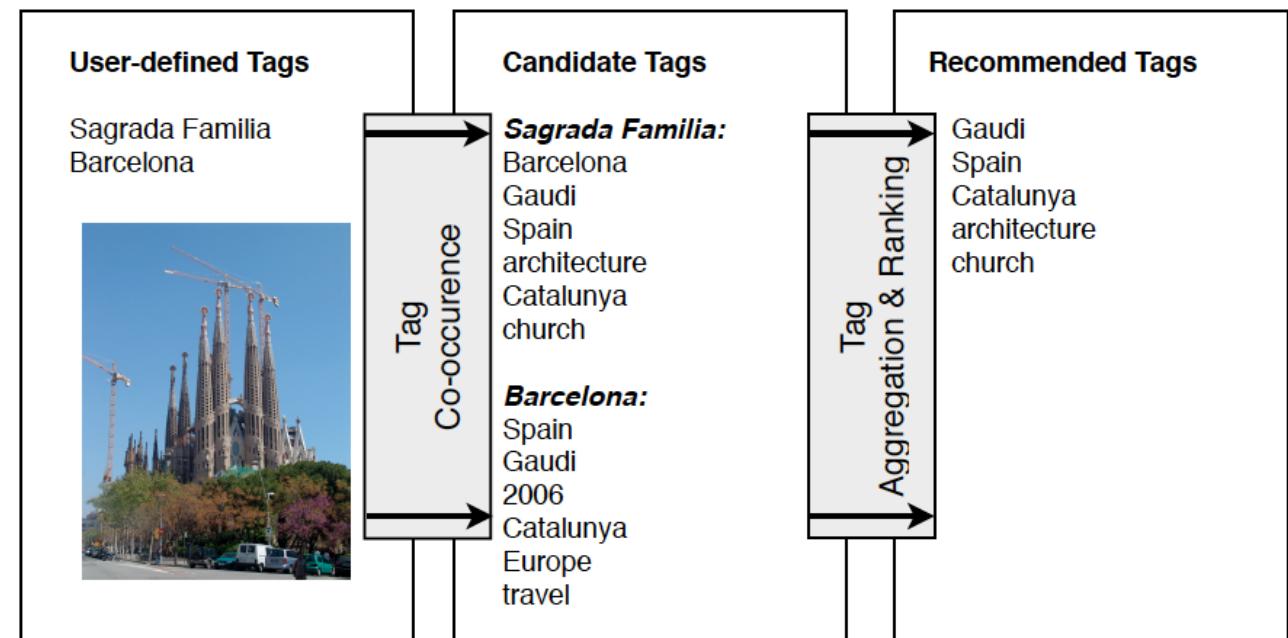
#### • 文本信息来源 (4) : 社交标注

- Web2.0时代，用户既是媒体服务的受益者，同时也是媒体数据的创造者。
  - 用户可以通过主动或被动的方式，为多媒体文件提供标注信息
    - 主动方式：为用户提供分享平台，鼓励用户对文档进行标注
    - 被动方式：以用户作为媒介，实现标签的有监督学习



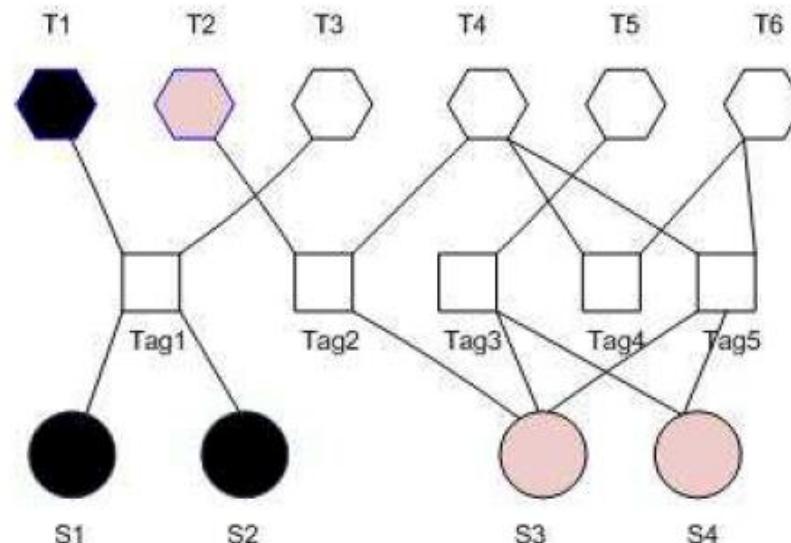
## • 文本信息来源 (4) : 社交标注

- 用户主动参与方式: 基于用户已有的部分标注, 补充更多的标签
- 核心思想: 基于标签之间的共现 (Co-occurrence)
  - 例如, 如果两个标签多次共同出现, 那么它们很可能可以同时用来描述一张新图片



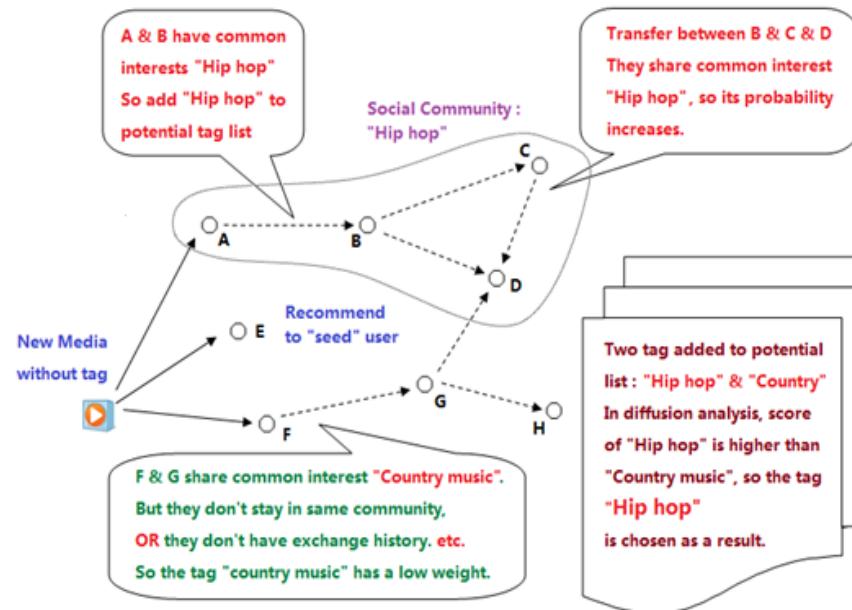
## • 文本信息来源 (4) : 社交标注

- 用户被动参与方式: 以用户为中介, 基于已有分类结果对其他实体分类
- 核心思想: 以已知的Tag  
(如Metadata) 为桥梁
- 通过图中的传播, 将已知实体的标签转移到未知标签的实体上



## • 文本信息来源 (4) : 社交标注

- 我们的发现：用户间的传播行为，可能揭示被传播实体的标签
- 核心思想：用户之间的分享行为往往存在主题敏感性
- 具有特定主题的小团体内的分享行为，往往意味着被分享的媒体文件具有相关主题



## • 基于文本的图像搜索：局限性

- 用户的需求难以用文字精确描述
  - 还是同样的问题：不愿意表达或者不知道如何表达
- 相比于文字，图像的需求更抽象，往往需要浏览更多文档才能发现和理解需求
- 更重要的是：多媒体文档往往难以用文字准确形容
  - 一图胜千言！



- 面向图像的检索
  - 基于图像内容的检索
  - 基于文本信息的检索
- 面向视频的检索
- 面向音频的检索
- 多模态混合检索

- 从图像到视频

- 视频是序列化的帧（图像）的集合，但又不是简单的图像串联
  - 视频包含更丰富的语义信息，如前后连续的动作、场景、剧情、人物关系等
  - 相比于语料标注积累日益丰富的图像数据集，视频数据更缺乏高质量的标注



## • 基本的视频检索：面向整段视频

- 面向整个视频的检索，往往基于视频元数据 / 标注 / 视频摘要进行

- 缺陷明显：标签可能无法涵盖视频全部内容，无法进行更细粒度的检索
  - 如：“马小帅入连仪式”在士兵突击第几集的第几分钟？



## • 解决方案 (1) : 逐帧分析

- 将视频视作帧（画面）的集合，进行帧或片段分析和检测
- 相比于单纯的图像分析，可以增加更多的特征
  - 主要是镜头相关特征，如镜头切换（直接、渐变）、运动（拉动、摇晃）
- 缺点：开支巨大，打散了视频连续的剧情等语义要素



## • 解决方案 (1) : 逐帧分析

- 重识别 (Re-identification) 可能有助于将离散的画面串联起来
  - 例如, 对监控画面中特定人物的重识别, 可以捕捉人物的连续行为轨迹
  - 然而, 除去计算开支大外, 这一技术在影视作品等视频中因假设不满足而难以运用
    - 人物重识别在镜头角度、人物姿态、衣着等方面均有一定要求



...

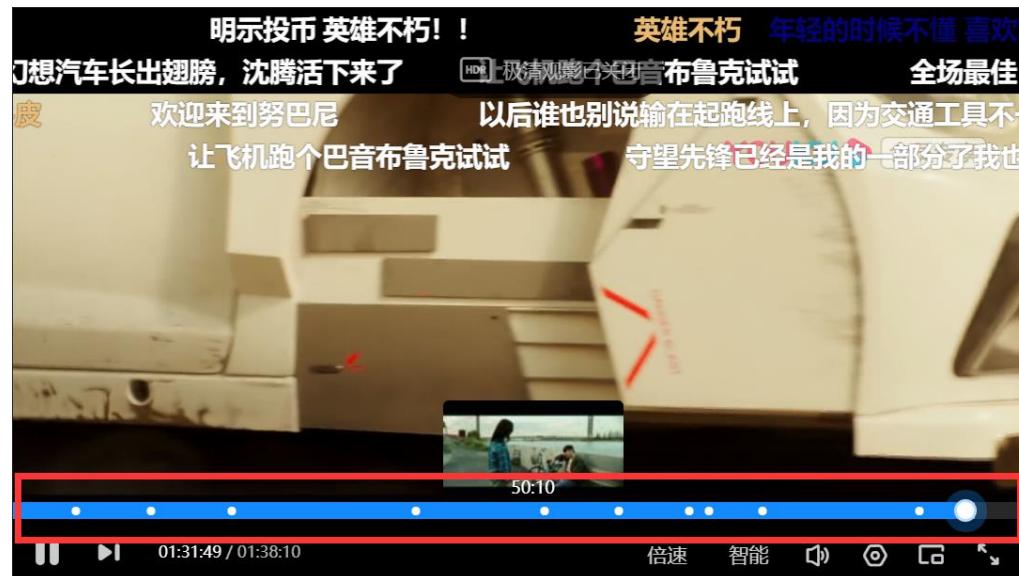


...



## • 解决方案 (2) : 重要片段提取

- 从视频中提取出最有意义的片段或者最有代表性的帧，并对这些内容进行标注
  - 用户对于整段视频各个部分的关注程度是不一样的，高潮部分最受关注
  - 提取最受关注的部分进行标注，有助于用户检索与快速浏览



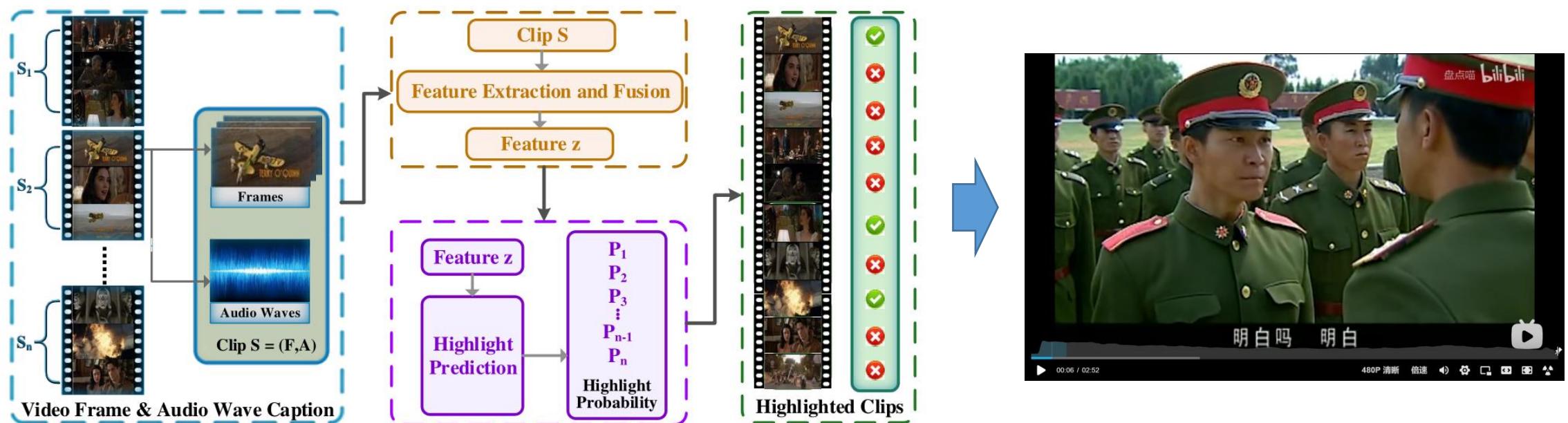
## • 解决方案 (2) : 重要片段提取

- 从视频中提取出最有意义的片段或者最有代表性的帧，并对这些内容进行标注
  - 一些启发式方法可以帮助我们获取代表性片段



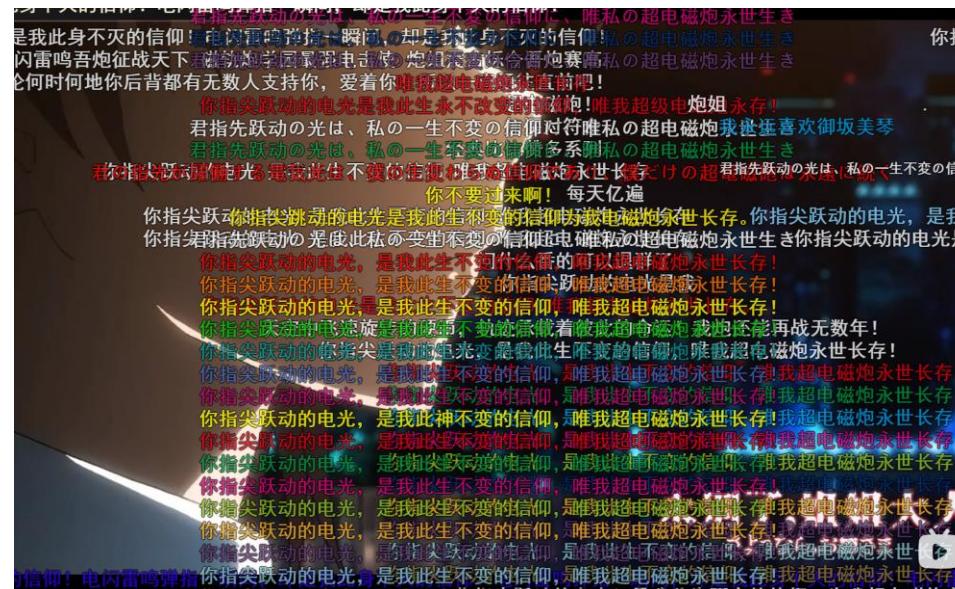
## • 解决方案 (2) : 重要片段提取

- 从视频中提取出最有意义的片段或者最有代表性的帧，并对这些内容进行标注
  - 从算法层面来看，模态中的一些语义线索也有助于挖掘重要片段
    - 例如，剧情高潮部分往往伴随着激昂的BGM或者主角高昂的语调



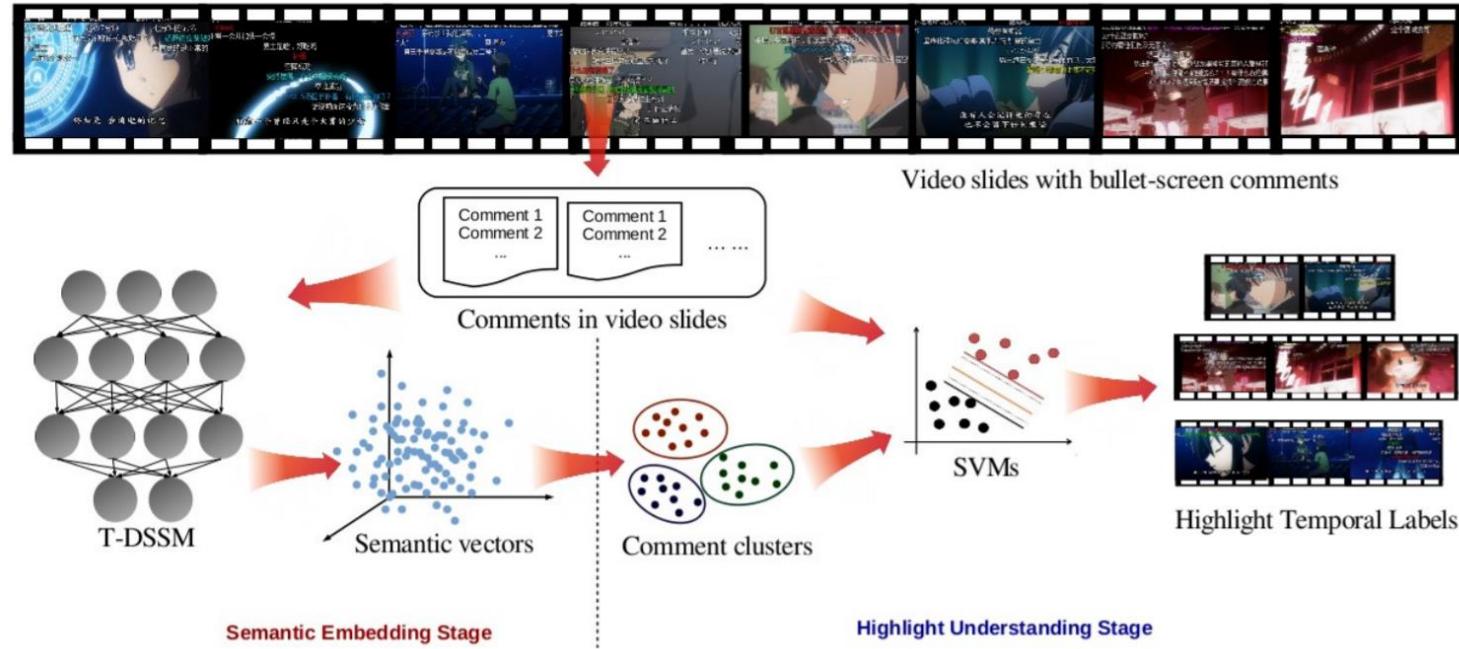
## • 解决方案 (3) : 视频的时序化标注

- 结合语义信息（如文本或标注），对视频各部分内容进行时序化标注
- 其难点在于缺乏足够语义标注，算法无法理解视频的高层次语义内容
- 新的线索：基于众包的语义标注信息，如弹幕



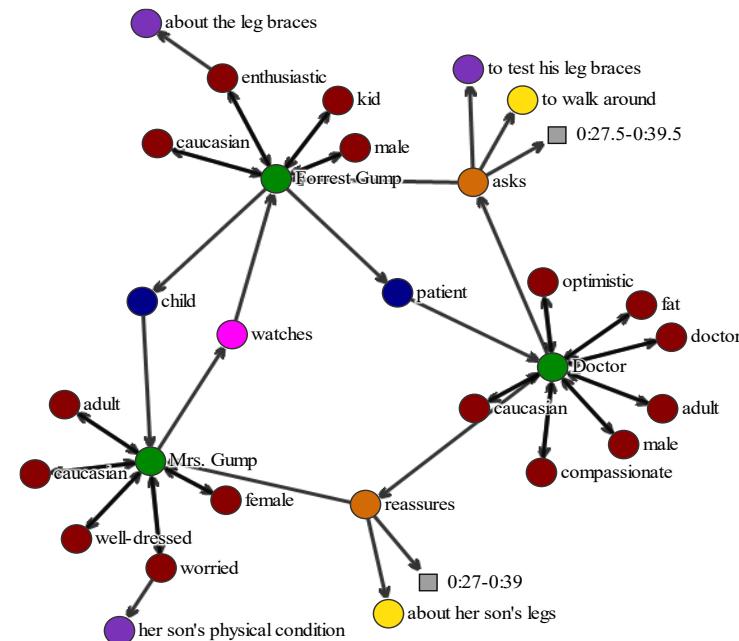
## • 解决方案 (3) : 视频的时序化标注

- 我们的尝试: 利用弹幕信息对视频各部分内容进行细粒度标注
  - 基本思路: 理解弹幕语义, 从中提炼若干主题, 并映射到相应视频片段上



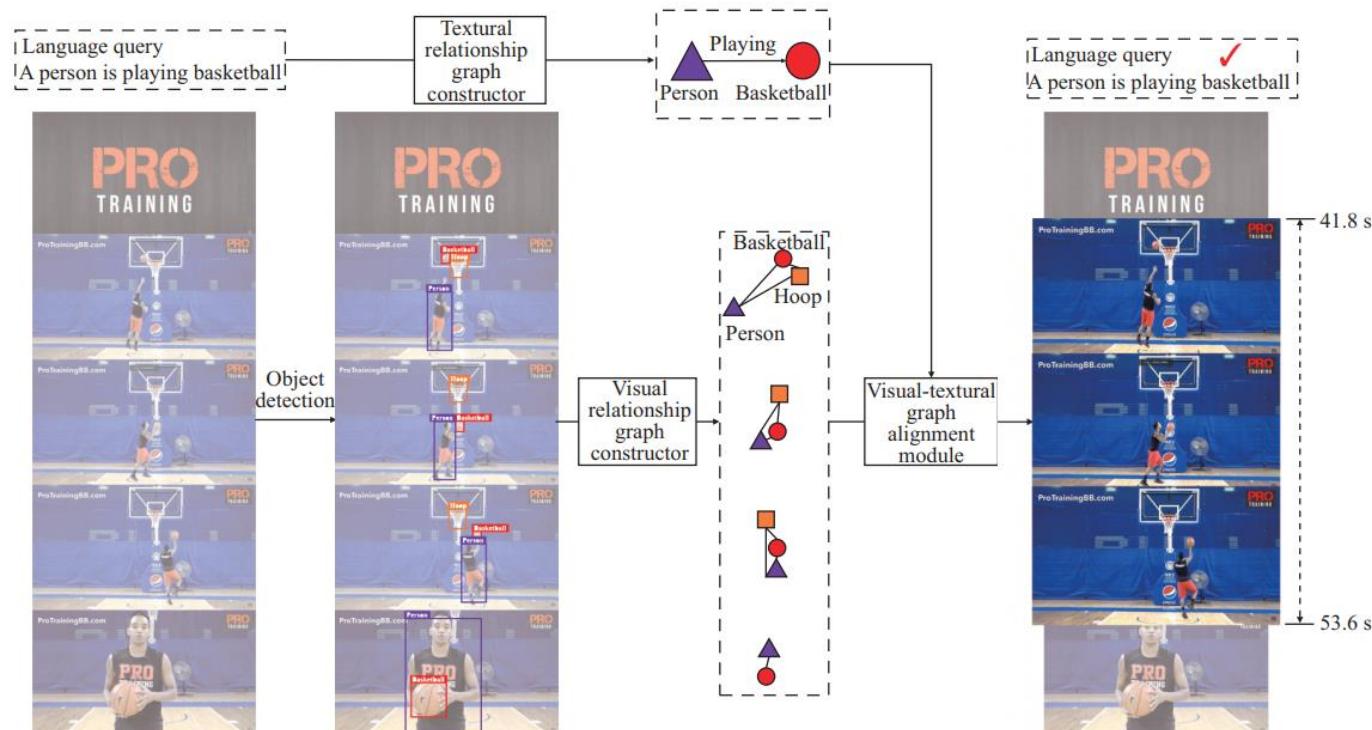
## • 解决方案 (4) : 基于场景图的细粒度视频描述

- 借助场景图 (Scene Graph) 技术, 实现视频内各种主体及其关系的描述
  - 然而, 此类方法受限于语义信息的获取难度, 在图结构的自动构建上仍存在诸多瓶颈



## • 解决方案 (4) : 基于场景图的细粒度视频描述

- 借助场景图 (Scene Graph) 技术, 实现视频内各种主体及其关系的描述
  - 通过引入文本信息, 构建语义化场景图, 建立从查询语句到细粒度视频帧的映射关系



基于视觉-文本关系对齐的跨模态视频片段检索, 陈卓等, 中国科学: 信息科学, 2020

- 面向图像的检索
  - 基于图像内容的检索
  - 基于文本信息的检索
- 面向视频的检索
- **面向音频的检索**
- 多模态混合检索

- 上古时代的音乐搜索
- 早先对于音频的搜索往往依赖于单纯的文字信息
  - 例如，对歌名、歌手等元信息或歌词进行搜索，容易因错漏而无法得到结果

钢铁锅,含眼泪喊修瓢锅 这是什么歌?  50

[我来答](#)

[分享](#)

[举报](#)

浏览 168546 次

39个回答



热心网友

2018-10-16

《海阔天空》

演唱: Beyond

#热议# 等的就是你! 有奖内测即将开始!

Baidu MP3 新闻 网页 贴吧 知道 MP3 图片 视频  
  
 铁频  歌词  全部音乐  mp3  rm  wma  其它格式  铃声  录音  
[把百度设为首页](#)  
 抱歉，没有找到与“稻香”相关的MP3内容。  
 百度建议您：  

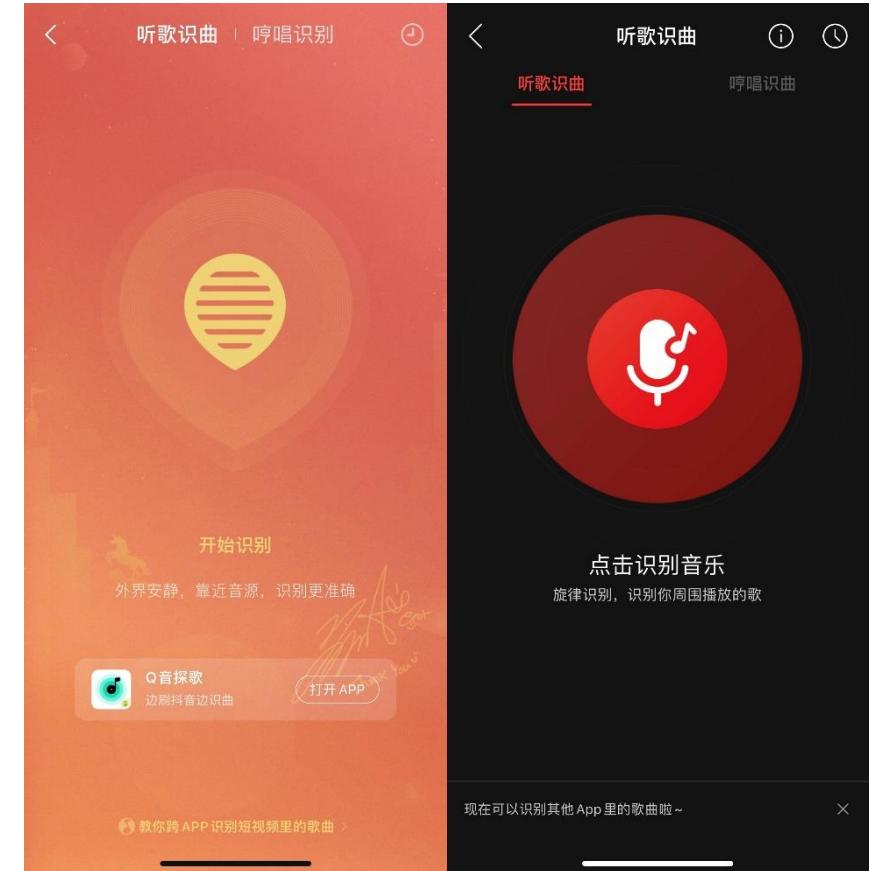
- 在百度网页中查找“稻香 mp3”
- 看看输入的文字是否有误
- 查看关于“稻香”的贴吧留言（3779篇）

 找到稻香相关视频约5814个 [查看更多>>](#)  


## 音频检索

- **从文字搜索到旋律搜索**

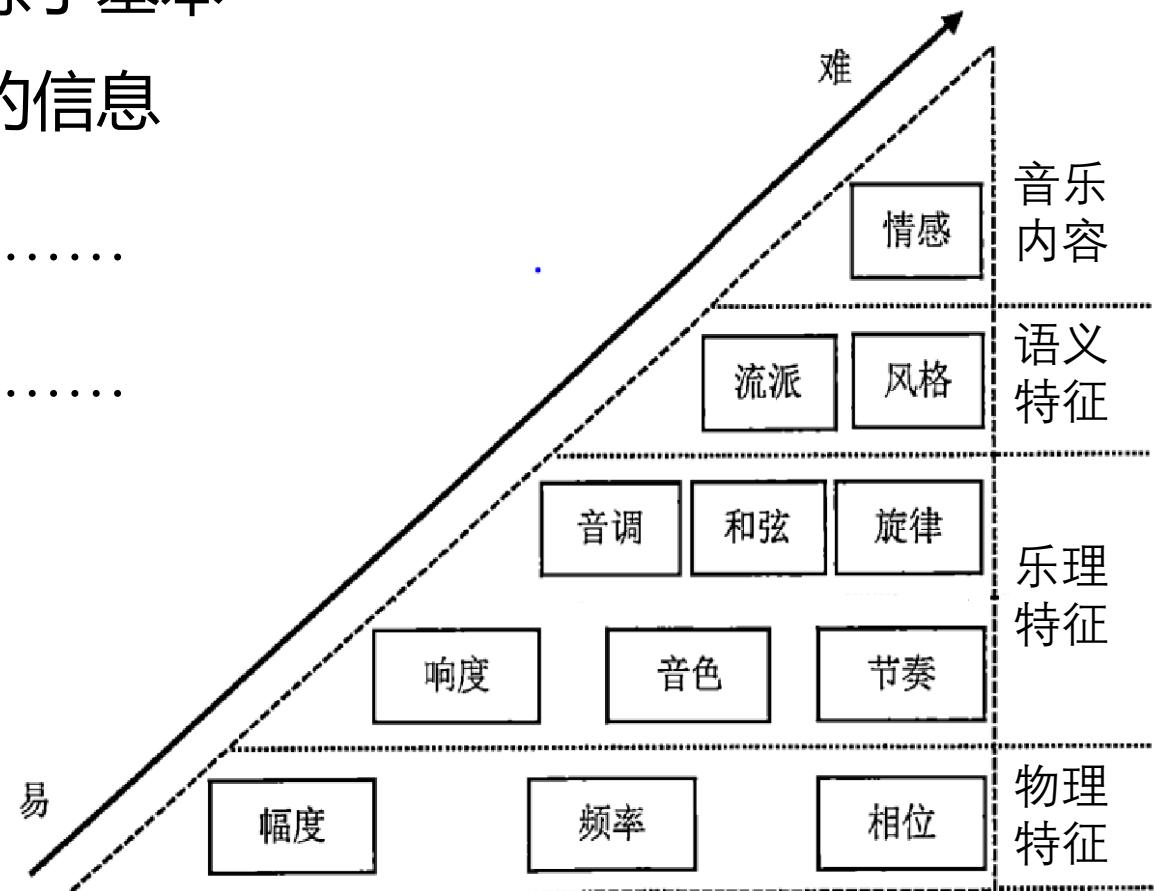
- 随着音频数据库的建立与丰富，以及音频匹配技术的发展，基于旋律搜索音乐已逐渐成为日常



## • 从文字搜索到旋律搜索

- 对于音频而言，其所包含的信息，除了基本的歌词语义外，还有许多其他种类的信息

- 音乐的种类：流行、摇滚、民谣.....
- 音乐的情感：悲伤、欢乐、激昂.....
- 音乐中蕴含的意象.....



## • 音乐类型 (1) : MIDI文件

- 乐器数字接口 (Musical Instrument Digital Interface)，编曲界最广泛的音乐标准格式，可称为“计算机能理解的乐谱”。
  - MIDI只能记录标准所规定的有限种乐器的组合，缺乏重现真实自然声音的能力，因此难以合成语音。



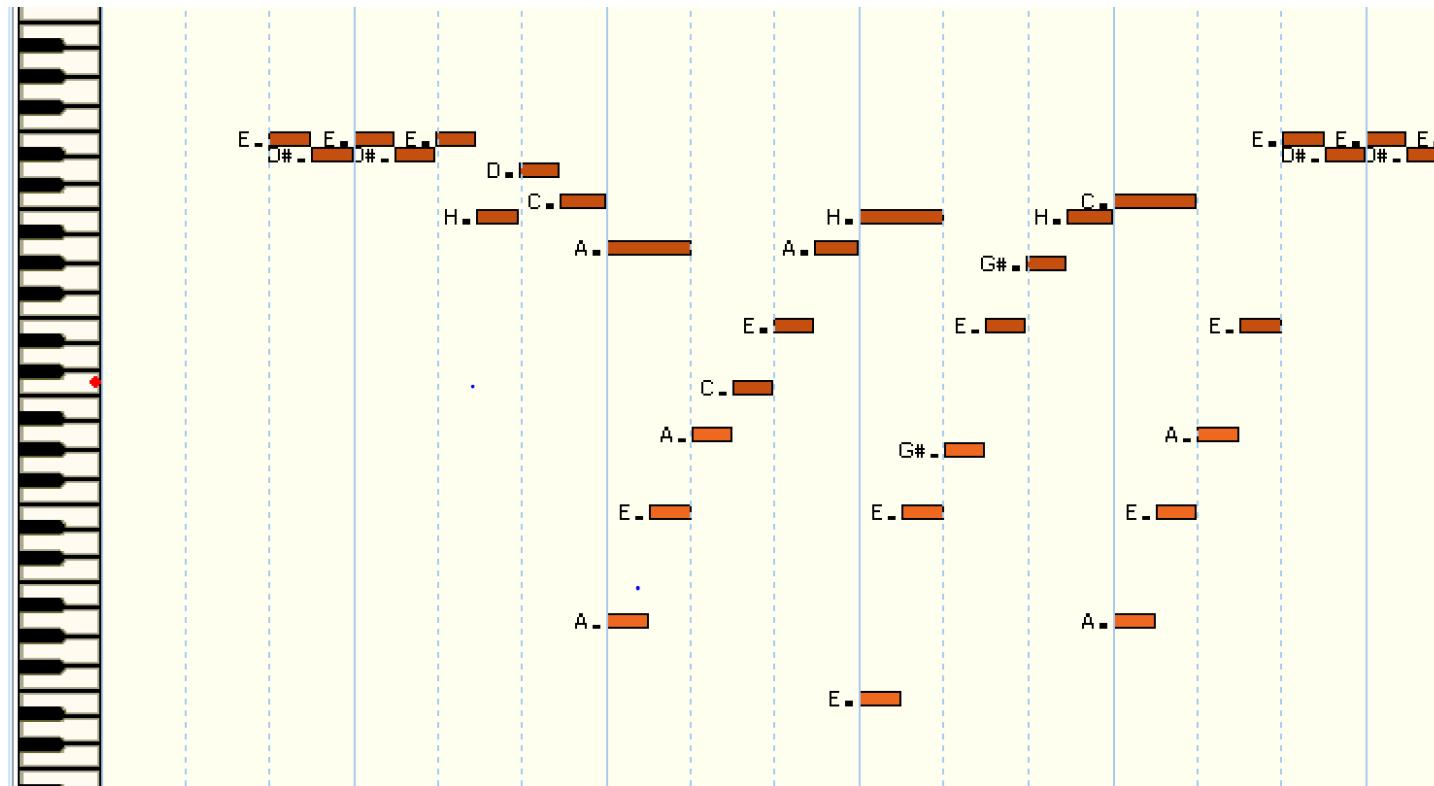
例子1



例子2 (带打击乐器)

## • 音乐类型 (1) : MIDI文件

- 下图是MIDI文件存储信息的可视化展示，由若干连续音符组成



## • 音乐类型 (1) : MIDI文件

- 对于音乐文件来说，主旋律是最具有代表性，容易被人感知的部分，且连贯性较好
- 对于MIDI文件，一般每个音轨对应一个通道，而主旋律的音符会有更大的演奏力度和更长的发声时间（对应MIDI文件中的“发声面积”，可以用这两者加以区别）
- 在得到主旋律音轨后，可以采用其旋律特征（音符序列）和节奏特征（时长）作为描述旋律的特征。

- 对于  $M$  中每个音轨  $T_i$ :
  - (1) 求  $T_i$  中音符的数量  $n$ , 同时发声的音符只计音调最高的一个, 下同;
  - (2) 记录音符序列  $P_i = (p_1, p_2, \dots, p_n)$ ,
  - (3) 记录音符力度  $V_i = (v_1, v_2, \dots, v_n)$ ,
  - (4) 记录音符起始时间序列  $S_i = (s_1, s_2, \dots, s_n)$ ,
  - (5) 计算音符的持续时长  $D_i = (d_1, d_2, \dots, d_n)$ ,
  - (6) 根据式 2.1 计算主旋律音轨度量指标  $\omega_i$ ,
- 选择  $\omega_i$  值最大的音轨  $T_\tau$ , 则:
  - (1)  $M$  的旋律特征:  $melody(M) = P_\tau$ ,
  - (2)  $M$  的节奏特征: 令  $(x_1, x_2, \dots, x_n) = S_\tau$  则:  
$$rhythm(M) = (x_1 - 0, x_2 - x_1, \dots, x_i - x_{i-1}, \dots, x_n - x_{n-1})$$
.

## • 音乐类型 (2) : 波形文件

- 波形文件包括MP3、 Audio、 Wave等大部分常用的音乐格式。使用范围相比MIDI文件更为广泛，也更适合记录音乐，但提取特征则相对更困难。
  - 难以分离不同的乐器声音和人声
  - 难以确定音符的音调
  - 难以确定音符的准确起始位置和长度
  - 容易受到噪声干扰

- 音乐类型 (2) : 波形文件

- 波形文件包括MP3、 Audio、 Wave等大部分常用的音乐格式。使用范围相比MIDI文件更为广泛，也更适合记录音乐，但提取特征则相对更困难。
  - 一个常见的任务场景：智能设备的语音控制

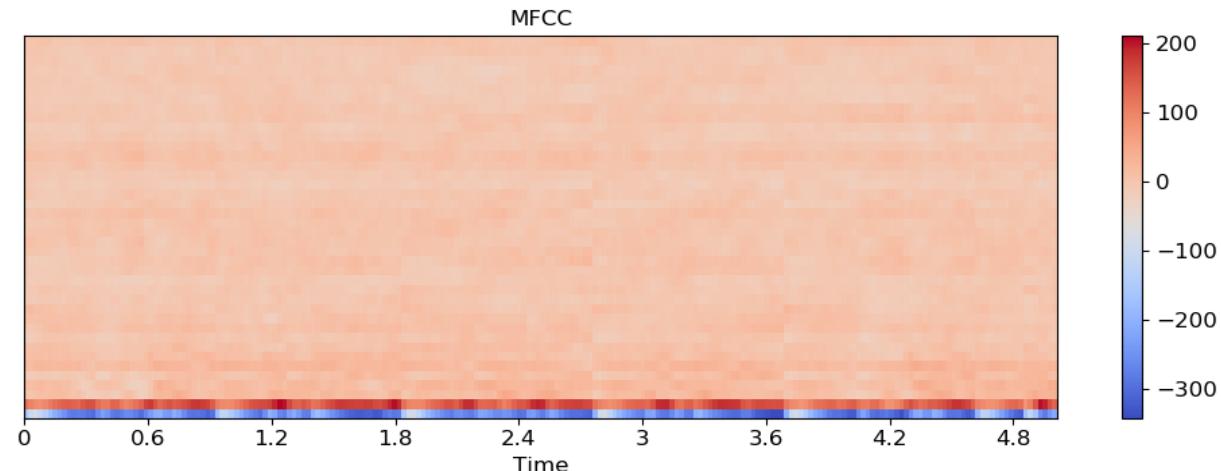


涉及技术点：语音识别、说话人识别, etc. →

- **音乐类型 (2) : 波形文件**

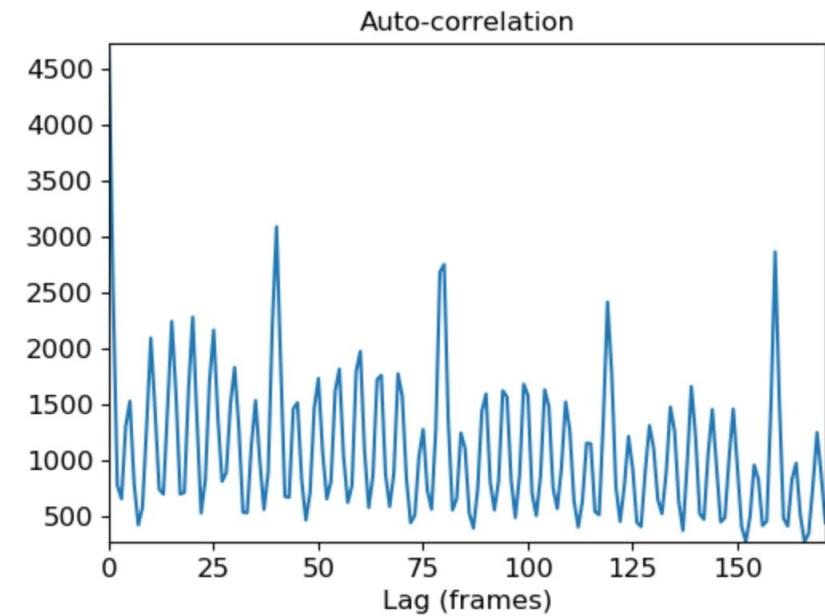
- 为描述音乐的音色特征，往往采用梅尔频率倒谱系数

- 梅尔频率倒谱系数 (Mel-frequency Cepstral Coefficients)，一种由模仿人类听觉得到的信号分析方法，在语音识别和音乐分类领域有着广泛应用。
- 在对信号预加重以加强信噪比的基础之上，求出信号的频谱特征，再经过若干转化，从而得到一种能够反映声源本身特点的信号特征。



## • 音乐类型 (2) : 波形文件

- 同时，还可采用提取基音序列的方式提取旋律特征
- 乐器或人发出的每一个乐音都是由不同频率和振幅的振动复合而成的，其中频率最低的振动发出的声音被称为基音，其余的被称为泛音，基音决定音高，因此提取基音序列即可视为提取旋律
- 采用自相关函数法（Autocorrelation Function）提取基因频率，自相关函数用于寻找信号中的重复模式，对信号进行自相关运算后，在基音周期的整数倍位置会出现峰值，因此寻找峰值的横坐标对应的频率即可推测基音频率。



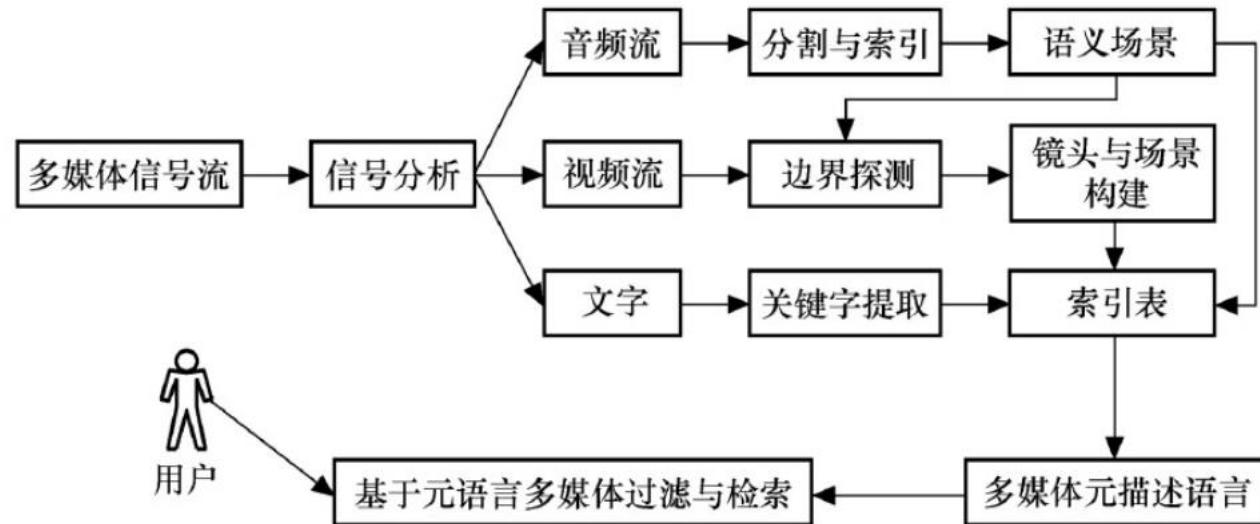
- 面向图像的检索
  - 基于图像内容的检索
  - 基于文本信息的检索
- 面向视频的检索
- 面向音频的检索
- 多模态混合检索

## • 多媒体融合分析与检索

- 先前介绍的各项工作，本质上仍以单媒体信息为主，多模态之间缺少融合
- 事实上，多媒体信息分析可借助各种媒体之间的关系融合进行。
  - 例如，分析一部电影的情节、人物关系和情感表达，显然要对文本、视频（行为、场景等）、音频信息进行综合分析，才能获得更好结果。
- 跨模态检索（Cross-modal Retrieval），指从非结构化的多媒体信息中检索出语义相似的同模态媒体，和语义相关的跨模态媒体的一种信息检索技术。

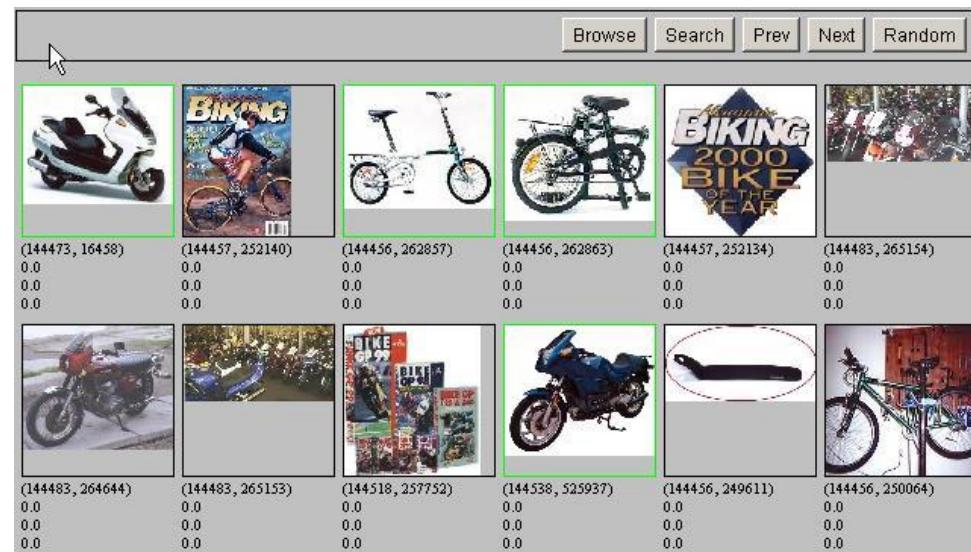
## • 不同融合策略（1）：单媒体交叉索引

- 首先对文字、音视频等单媒体信息分别处理。
- 在此基础之上，用生成结果对自身或其他媒体数据流进行索引。
  - 例如，音频中的连续对话有助于帮助判断一个连续的场景，进而协助视频的边界探测



## • 不同融合策略（2）：单媒体结果融合

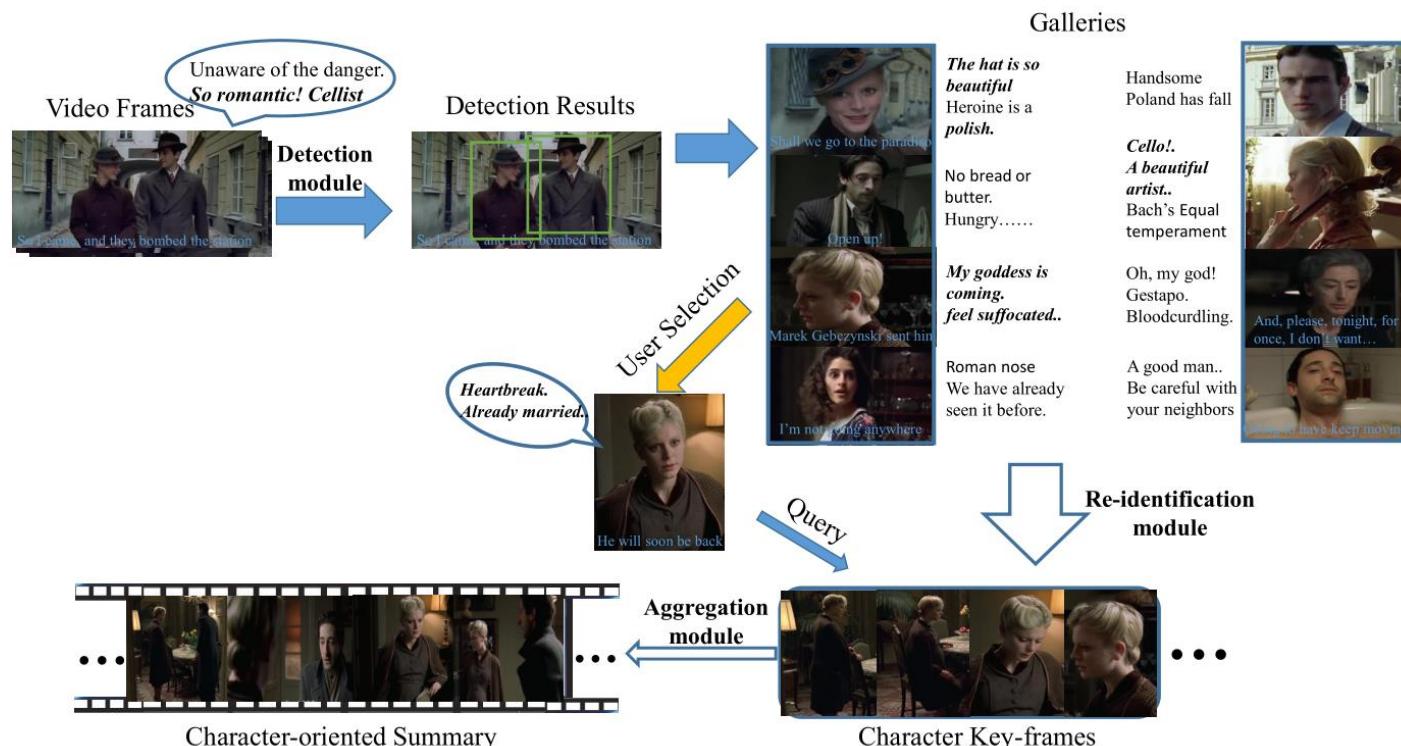
- 首先得到基于单媒体的检索结果，再将不同媒体的结果融合起来。
- 相当于每种单媒体作为一种感知器，得到一组检索结果，进行融合决策
  - 例如，先前用户相关性反馈的“Bike”的例子，就是文本与图像结果的融合



没错还是这张图

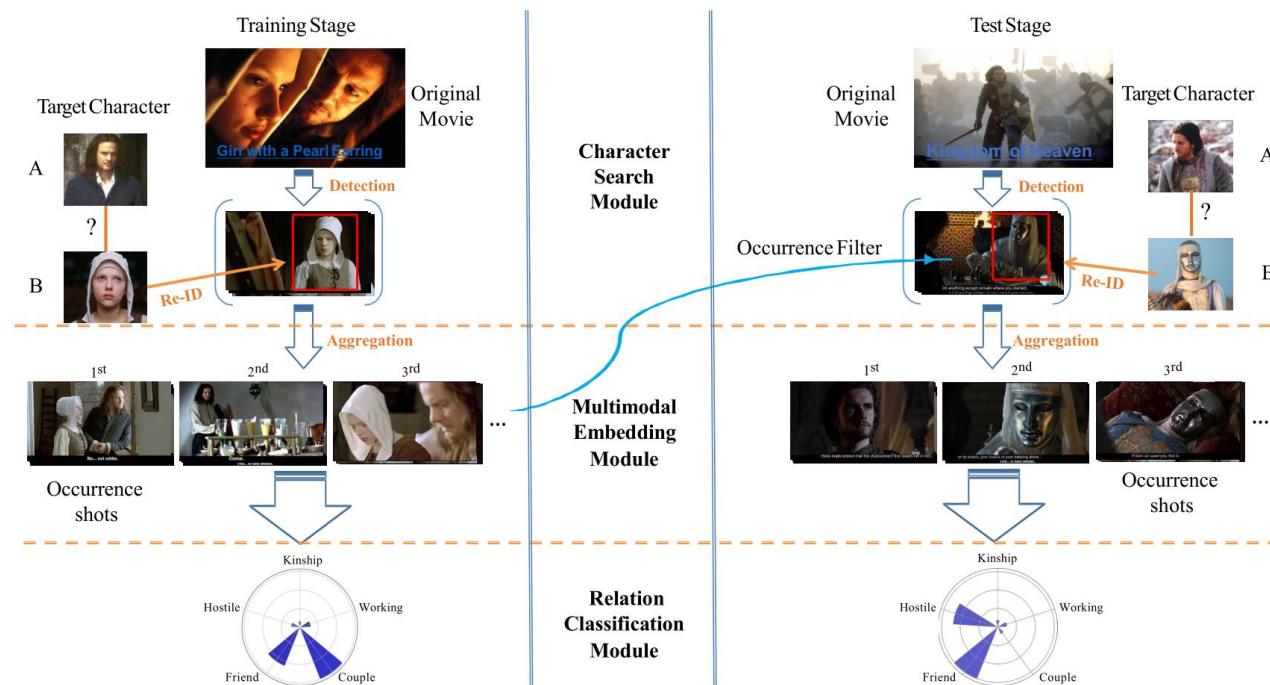
## • 不同融合策略（3）：多媒体特征融合

- 在多媒体语义对齐的基础之上，实现对于媒体信息的跨模态联合建模。
- 我们的尝试：图文联合建模，借助文本信息协助寻找特定人物出场的视频片段



## • 不同融合策略 (3) : 多媒体特征融合

- 在多媒体语义对齐的基础之上，实现对于媒体信息的跨模态联合建模。
  - 我们的尝试：图文联合建模，借助文本信息协助判断特定人物之间的关系



# 本章小结

## 多模态检索

- 面向图像的检索
  - 基于内容/基于文本信息的检索方法规则
- 面向视频的检索，结合时序文本的图像分析
- 面向音频的检索，Midi/波形的不同特征
- 跨模态分析
  - 三种不同类型的模态融合，跨模态信息对齐

[tongxu@ustc.edu.cn](mailto:tongxu@ustc.edu.cn)