图象工程（下）

图 象 理 解

（第4版）

章毓晋
清华大学电子工程系 100084 北京

第4单元　研究示例

- 第13章　多传感器图象信息融合
- 第14章　基于内容的图象和视频检索
- 第15章　时空行为理解

一些得到较多关注的研究领域

结合利用不同传感器所获得的数据

检索是各类视觉信息在全球得到广泛采集、传输和应用背景下一个新的研究领域

图象理解需要充分掌握时空信息，分析人物行为，解释场景含义

第14章　基于内容的图象和视频检索

14.1　图象和视频检索原理

视觉信息检索

从视觉数据库（集合）中快速地提取出与一个查询相关的图象集合或图象序列

1、基于内容的检索

基于内容的视觉信息检索（CBVIR）
基于内容的图象检索（CBIR）
基于内容的视频检索（CBVR）

多媒体内容描述界面（MPEG-7）

14.1　图象和视频检索原理

2、归档和检索流程图

图象归档：提取图象或目标的视觉特征

图象检索：采用范例查询方式，对给定的查询图，进行相应的分析并提取其匹配特征进行检索

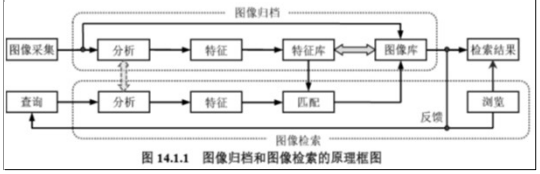


图 14.1.1　图像归档和图像检索的原理框图

14.1　图象和视频检索原理

2、归档和检索流程图

三个关键：选择恰当的图象特征，有效的特征提取方法，准确的特征匹配算法

五个功能模块：提供查询手段，描述图象内容，匹配图象内容，提取对应图象，验证检索结果

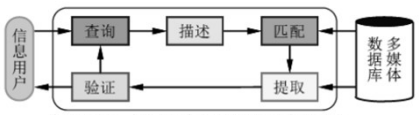


图 14.1.2　图像检索系统的 5 个功能模块

14.1 图象和视频检索原理

3、多层次内容表达

人类在对图象内容进行描述时，常使用语义层次的概念和术语

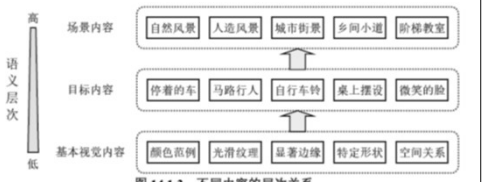


图 14.1.3 不同内容的层次关系

14.1 图象和视频检索原理

3、多层次内容表达

语义除可以描述客观事物外，还可以描述主观感受以及更抽象的概念（如广泛、富有等）

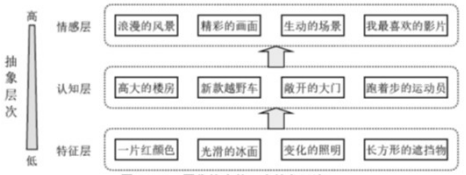


图 14.1.4 图像检索的三个抽象层次

14.2 视觉特征的匹配和检索

按范例查询，把用作参考的查询图象与图象库中的大量图象借助视觉特征进行匹配

14.2.1 颜色特征匹配

- **颜色特征的统计直方图**（上册）

1、直方图相交法

$$P(Q,D)=\frac{\sum_{k=0}^{L-1}\min[H_Q(k),H_D(k)]}{\sum_{k=0}^{L-1}H_Q(k)}$$

2、距离法

$$\boldsymbol{f}=[\ \mu_R \quad \mu_G \quad \mu_B\ ]^{\mathrm{T}}$$

$$P(Q,D)=\sqrt{(\boldsymbol{f}_Q-\boldsymbol{f}_D)^2}=\sqrt{\textstyle\sum_{R,G,B}(\mu_Q-\mu_D)^2}$$

14.2.1 颜色特征匹配

3、 中心矩法

更高阶的矩（三阶）

$$P(Q,D)=\sqrt{W_R\sum_{i=1}^{3}(M_{QR}^i-M_{DR}^i)^2+W_G\sum_{i=1}^{3}(M_{QG}^i-M_{DG}^i)^2+W_B\sum_{i=1}^{3}(M_{QB}^i-M_{DB}^i)^2}$$

4、 参考颜色表示法

将图象颜色用一组参考色表示

$$\boldsymbol{f}=[r_1 \quad r_2 \quad \cdots \quad r_N]^{\mathrm{T}}$$

$$P(Q,D)=W\sqrt{(\boldsymbol{f}_Q-\boldsymbol{f}_D)^2}=\sqrt{\sum_{i=1}^{N}W_i(r_{iQ}-r_{iD})^2}$$

14.2.2 纹理特征计算

- 统计法是纹理特征提取的一类有效方法
  常以灰度级空间相关矩阵为基础
- **灰度共生矩阵**$M_{(\Delta x,\Delta y)}(h,k)=m_{hk}$（中册）

(1) 反差 $G=\sum_h\sum_k(h-k)^2m_{hk}$

(2) 能量 $J=\sum_h\sum_k(m_{hk})^2$

(3) 熵 $S=-\sum_h\sum_k m_{hk}\log m_{hk}$

(4) 相关 $C=\dfrac{\sum_h\sum_k hkm_{hk}-\mu_x\mu_y}{\sigma_x\sigma_y}$

14.2.2 纹理特征计算

- 结合采用各个参数的均值和标准差
- 纹理特征向量中的各个分量

$$\mu_G, \sigma_G, \mu_J, \sigma_J, \mu_S, \sigma_S, \mu_C, \sigma_C$$

- **高斯归一化**

特征向量 $\boldsymbol{F}_i = [f_{i,1}\ f_{i,2}\ \cdots f_{i,N}]$

将$f_{i,j}$归一化至[–1, 1]区间

各个$f_{i,j}$均转变成具有$N(0, 1)$分布

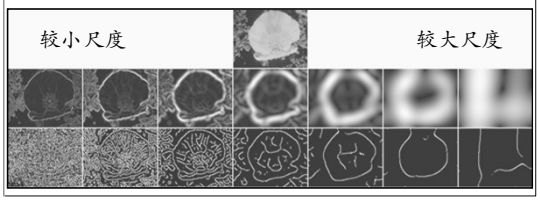
$$f_{i,j}^{(N)} = \frac{f_{i,j} - m_j}{\sigma_j}$$

14.2.3 多尺度形状特征

- （目标）形状特征可以看作是比（象素）颜色或邻域纹理要更高层次的特征
- 目标的形状常可用目标的轮廓来表示
- 轮廓是由一系列边界点所组成的
- 在较大尺度下常能较可靠地消除误检并检测到真正的边界点，但在大尺度下对边界的定位不易准确
- 在较小尺度下对真正边界点的定位常比较准确，但在小尺度下误检的比例会增加

14.2.3 多尺度形状特征

- **小波变换模极大值**（wavelet modulus maxima）
- 先在较大尺度下检测出真正的边界点，再在较小尺度下对真正边界点进行较精确的定位



14.2.4 综合特征检索

- 结合使用多个单一特征

特征\库	Holidays (mAP%)	Oxford (mAP%)	Paris (mAP%)	Flickr100K (mAP%)	UKBench (N-S)
HSV	61.95	—	—	53.38	3.195
SIFT	80.74	**76.66**	75.98	72.06	3.504
CNN	71.67	43.28	63.64	62.26	3.490
融合	**93.72**	75.56	**90.61**	**90.91**	**3.939**

mAP：与多个查询图象对应结果的AP值的平均

N-S值：前4幅查询结果的平均召回数量

14.3 基于运动特征的视频检索

- 运动信息表示视频内容沿时间轴的发展变化
- 运动特征是视频数据所独有的特征
- 视频序列中的运动信息可分为两类

14.3.1 全局运动特征

14.3.2 局部运动特征

14.3.1 全局运动特征

- 摄象机运动所造成的全部点的整体移动
- 用短时运动分析法提取相邻帧间的运动信息
- 为了得到有意义的运动内容，需要将各个短时运动分析的结果按时间顺序结合起来
- **特征点序列：**

将运动信息用运动特征空间中的一个点来表示，而把一个较长时间的运动表示为运动特征空间中的一个点序列

14.3.1 全局运动特征

- 对特征点序列进行相似性度量可借用字符串匹配的方法
- 两个特征点序列l_1和l_2可分别表示为

$$\{f_1(i),\ i=1,2,\ldots,N_1\}$$

$$\{f_2(j),\ j=1,2,\ldots,N_2\}$$

✓ 序列的长度相同，即$N_1=N_2=N$

$$S(l_1,l_2)=\sum_{i=1}^{N}S_f[f_1(i),f_2(i)]$$

14.3.1 全局运动特征

✓ 如果序列的长度不同，设$N_1 < N_2$

- 在l_2中以不同的时间起点t截取与l_1长度相同的序列$l'_2(t)$
- 通过移动时间起点t，还可以计算出对应所有可能的时间起点t的子序列的相似度。而两个序列l_1和l_2的相似度可选为其中的最大值

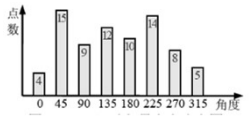
$$S(l_1,l_2)=\max_{0\le t\le N_1-N_2}\sum_{i=1}^{N}S_f[f_1(i),f_2(i+t)]$$

14.3.2 局部运动特征

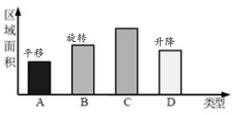
局部运动特征匹配

- 运动体现在对应场景中运动目标的位置
- 运动的表现比较复杂，不太规律
- 借助局部运动矢量场表达运动信息

运动矢量的方向直方图



运动区域的类型直方图



14.4 视频节目分析和索引

视频分析的目的是建立/恢复视频中的（语义）结构，并根据这个结构进行查询或建立索引

14.4.1 新闻视频结构化

14.4.2 体育比赛视频排序

14.4.3 家庭录象视频组织

14.4.1 新闻视频结构化

1. 新闻节目视频的特点

- 结构特征比较明显
- 各种类型的新闻节目具有较好的一致性
- 主体内容均是由一系列新闻故事单元组成
- 每一个新闻故事单元，即一个新闻条目，讲述一个在内容上相对独立的事件，具有明确的语义。它是新闻节目视频分析、索引和查询的基本单位

14.4.1 新闻视频结构化

2. 重要说话人镜头检测

- 新闻画面中出现的单个人物说话头像的近镜头（视频画面中主要对象为人物头像）
- 说话人镜头的时间、空间和运动的视觉特征：

(1) 画面的整体运动变化比一般的镜头小很多，但是却要比那种拍摄几乎静止画面的镜头大

(2) 画面运动主要集中于一个固定的说话人头部而且这个头部的位置在新闻视频中集中于相对固定的三个位置：中部，左侧和右侧

14.4.1 新闻视频结构化

3. 重要说话人镜头聚类

- 按照人物进行聚类
  - 首先提取镜头的颜色特征
  - 然后采用彩色直方图相交计算镜头的相似度
  - 利用位置模型可以更准确地计算相似度
  - 采用无监督聚类的方法对所有的重要说话人镜头聚类
- 利用对新闻标题条的检测结果来去除非重要人物（即误检）的镜头类

14.4.2 体育比赛视频排序

1. 体育比赛视频的特点

- 体育比赛节目一般也有较强的结构性
- 体育比赛进行的环境是特定的
- 体育比赛中有许多不定因素（时间、位置）
- 体育比赛本身总有一些高潮事件
- 体育比赛节目是（基于）事件的视频
- 使用先验知识对精彩事件进行定义，并通过检测体育比赛中的特定事件来完成对精彩镜头的检测（对镜头排序）

14.4.2 体育比赛视频排序

2. 乒乓球比赛节目的结构

- 乒乓球比赛基于比分（时长不固定）
- 比赛节目分为：发球事件、比赛事件、场间休息、观众场面以及回放片段

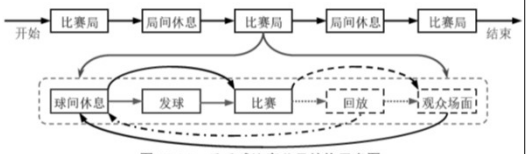


图 14.4.4 乒乓球比赛节目结构示意图

14.4.2 体育比赛视频排序

3. 目标检测和跟踪

- 要统计乒乓球比赛的客观指标，需要先对场景中的目标进行检测，包括运动员位置检测、球桌位置检测和球位置检测

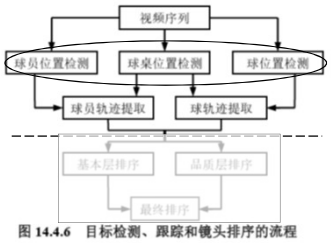


图 14.4.6 目标检测、跟踪和镜头排序的流程

14.4.2 体育比赛视频排序

4. 精彩度排序

分两个层次（物理特征和抽象概念）

(1) 基本层的排序

- 排序指标

$$R = N\left(w_{\mathrm{v}}h_{\mathrm{v}} + w_{\mathrm{b}}h_{\mathrm{b}} + w_{\mathrm{p}}h_{\mathrm{p}}\right)$$

① 球运动的平均速度

$$h_{\mathrm{v}} = f\left(\sum_{i=1}^{N}|v(i)|\Big/N\right)$$

Sigmoid函数

$$f(x) = \frac{1}{1+\exp[-(x-\bar{x})]}$$

14.4.2 体育比赛视频排序

4. 精彩度排序

(1) 基本层的排序

② 连续两次击球之间球运动的平均距离

$$h_{\mathrm{b}} = f\left(\sum_{i=1}^{N_1}|b_1(i+1)-b_1(i)|\Big/N_1 + \sum_{i=1}^{N_2}|b_2(i+1)-b_2(i)|\Big/N_2\right)$$

③ 连续两次击球之间运动员运动的平均距离

$$h_{\mathrm{p}} = f\left(\sum_{i=1}^{N_1}|p_1(i+1)-p_1(i)|\Big/N_1 + \sum_{i=1}^{N_2}|p_2(i+1)-p_2(i)|\Big/N_2\right)$$

14.4.2 体育比赛视频排序

4. 精彩度排序

(2) 品质层的排序（借助一些高层的概念）

- 运动员移动的激烈程度（位置和形状加速度）

$$m(i)=w_{\mathrm{p}}f\left[\left|p(i)-p(i-2)\right|\right]+w_{\mathrm{s}}f\left[\left|s(i)-s(i-2)\right|\right]$$

- 球轨迹的品质（轨迹长度和瞬时速度）

$$t(i)=w_{\mathrm{l}}f\left[l(i)\right]+w_{\mathrm{v}}f\left[v(i)\right]$$

- 击球的变化（不相似性，d对应方向）

$$u(i)=w_{\mathrm{v}}f\left[v(i)-v(1-i)\right]+w_{\mathrm{d}}f\left[d(i)-d(1-i)\right]+w_{\mathrm{l}}f\left[l(i)-l(1-i)\right]$$

14.4.3 家庭录象视频组织

- ✓ 家庭录象主要记录了人们的生活，而不是讲述人工编造的故事，所以其中的各个镜头常是同样重要的
- ✓ 家庭录象主要由没有编辑过的原始视频镜头组成，是按时间标记的数据
- ✓ 家庭录象视频的拍摄目的不是为了广大观众（如广播视频那样），而是为了朋友、客人和家人。所以对其的分析应该考虑尽量体现拍摄者的拍摄手法和意图

14.4.3 家庭录象视频组织

1. 运动关注区域检测

- 注意力集中在与画面整体运动不同（即有明显的相对运动）的区域中，这些区域称为运动关注区域
- 其他相对静止的区域给人形成环境的概念，它表达了视频内容给人造成的总体印象

(a) (b) (c)

图 14.4.9 视频画面的空间分割

14.4.3 家庭录象视频组织

2. 基于摄象机运动的时间加权模型

- 4参数运动模型

$$\begin{cases} u=h_0x+h_1y+h_2 \\ v=h_1x+h_0y+h_3 \end{cases}$$

- 镜头运动参数 ⇓

帧间镜头缩放 $S=h_0+1$

绕Z轴旋转角度 $r=h_1/(h_0+1)$

摇镜头的幅度 $L=\sqrt{(a\lambda)^2+(\lambda b)^2}=\sqrt{h_2^2+h_3^2}\,/(h_0+1)$

摇镜头的角度 $\theta=\arctan[(b\lambda)/(a\lambda)]=-\arctan(h_3/h_2)$

14.4.3 家庭录象视频组织

2. 基于摄象机运动的时间加权模型

- ✓ 将摄象机运动参数映射到视觉关注度
- ✓ 建模的主要假设为：
  - 镜头缩放的作用是强调，镜头放大用于强调画面局部细节，镜头缩小用于强调画面整体印象
  - 摇镜头的情况需要分为目标跟踪（对应有运动关注区域的情况）和环境转换（对应无运动关注区域的情况）两种类型
  - 在水平或垂直摇镜头的运动十分频繁变化而且运动幅度较小时，它将被认为是随机的不稳定抖动

14.4.3 家庭录象视频组织

3. 镜头组织策略

- 两级的镜头组织策略

(1) 第一级镜头聚类层次检测出场景的变换，即运动区域和环境特征同时发生较大变化的位置，它代表情节和地点的转移

(2) 第二级镜头聚类层次在一个场景内部进一步检测运动区域或环境特征中某一方面发生变化（不包括同时变化）的情况，它代表了在同一镜头内：

- 关注焦点的变化（不同的运动目标，相同的环境）
- 关注目标的位置转移（相同的运动目标，不同的环境）

14.4.3 家庭录象视频组织

3. 镜头组织策略

- 第一级镜头聚类将5个镜头分为3个场景单元
- 第3和第4画面放在一起作为第二级镜头聚类

图 14.4.12 两级镜头聚类

亚场景

14.5 语义分类检索

视觉特征 ⇒⇒⇒⇒⇒ 语义特征

特征层 ⇒⇒ 认知层 ⇒⇒ 情感层
（客观） （主观）

14.5.1 基于视觉关键词的图象分类

14.5.2 高层语义与气氛

14.5.2 高层语义与气氛

从人的认知角度看，人对图象的描述和理解主要是在语义层次进行的

语义除可以描述客观事物（如图象、教室、摄象机等）外，还可以描述主观感受（如漂亮、清晰等）以及更抽象的概念（如广泛、富有等）

除了有**认知水平**的目标语义、场景语义等，还有主观性更强的**抽象属性**语义（如气氛、情感等）

14.5.2 高层语义与气氛

1. 五种气氛语义

- 利用全局照度（照度分布）和主要色调这两个特征的不同组合可定义5种典型的气氛

表 14.5.1 5种气氛的照度和色调特点

编号	气氛	照度（对比度）	色调
1	有活力和强劲（vigor and strength）	照度大，对比度大	鲜艳
2	神秘或恐怖（mystery or ghastfulness）	对比度大	黯淡/幽深
3	兴奋和明亮（victory and brightness）	照度大，对比度小	暖色调
4	平静或凄惨（peace or desolation）	对比度小	冷色调
5	不协调（lack unity）和离析（disjoint）	分布零乱	—

14.5.2 高层语义与气氛

2. 分级分类

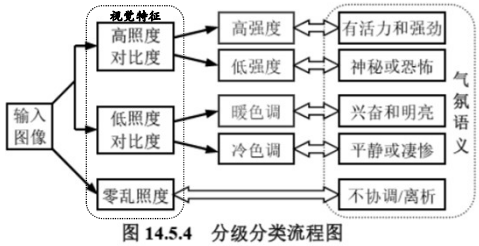


图 14.5.4 分级分类流程图

联 系 信 息

- 通信地址：北京清华大学电子工程系
- 邮政编码：100084
- 办公地址：清华大学，罗姆楼，6层305室
- 办公电话：(010) 62798540
- 传真号码：(010) 62770317
- 电子邮件：zhang-yj@tsinghua.edu.cn
- 个人主页：oa.ee.tsinghua.edu.cn/~zhangyujin/