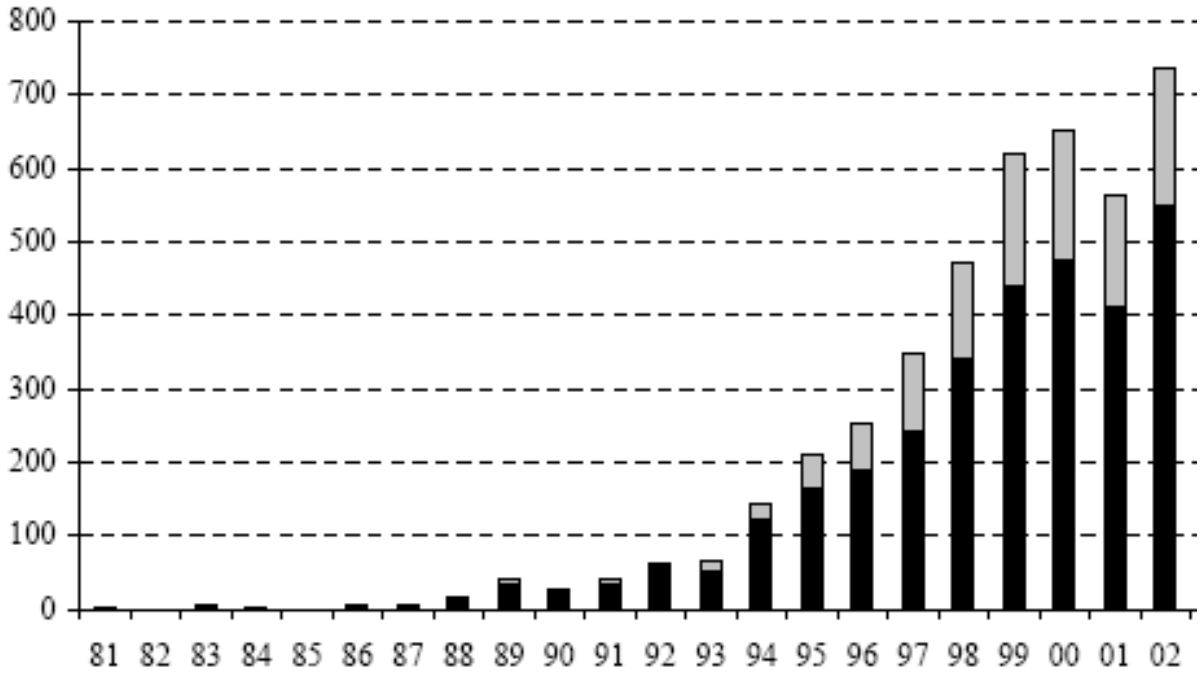
视频分析与检索技术

Video Analysis and Retrieval Technology

Active Research Area



- Graph from “A new perspective on Visual Information Retrieval”, Horst Eidenberger, 2004
- Black: “Image Retrieval”; Grey:”Video Retrieval”; IEEE Digital Library

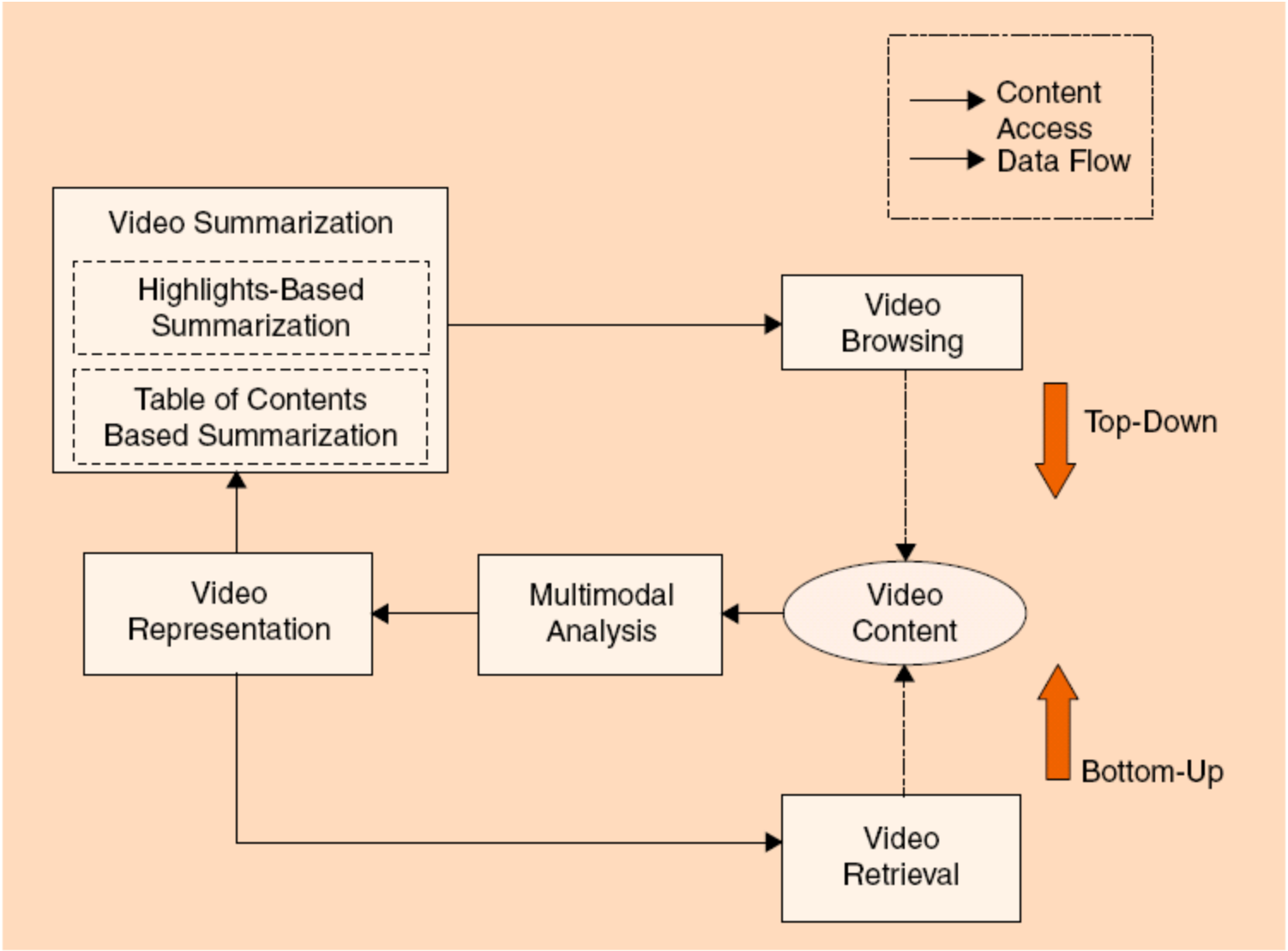
- TRECVID 2003 **Event**
  - "Find shots of an airplane taking off."

- TRECVID 2004 **Named-person Location**
  - "Find shots of Bill Clinton speaking with a US flag visible behind him."

- IBM Speech Group **Objects**
  - "Find shots containing monkeys or gorillas."

- BBC Logs **Named-location**
  - "Find shots of the Kremlin."

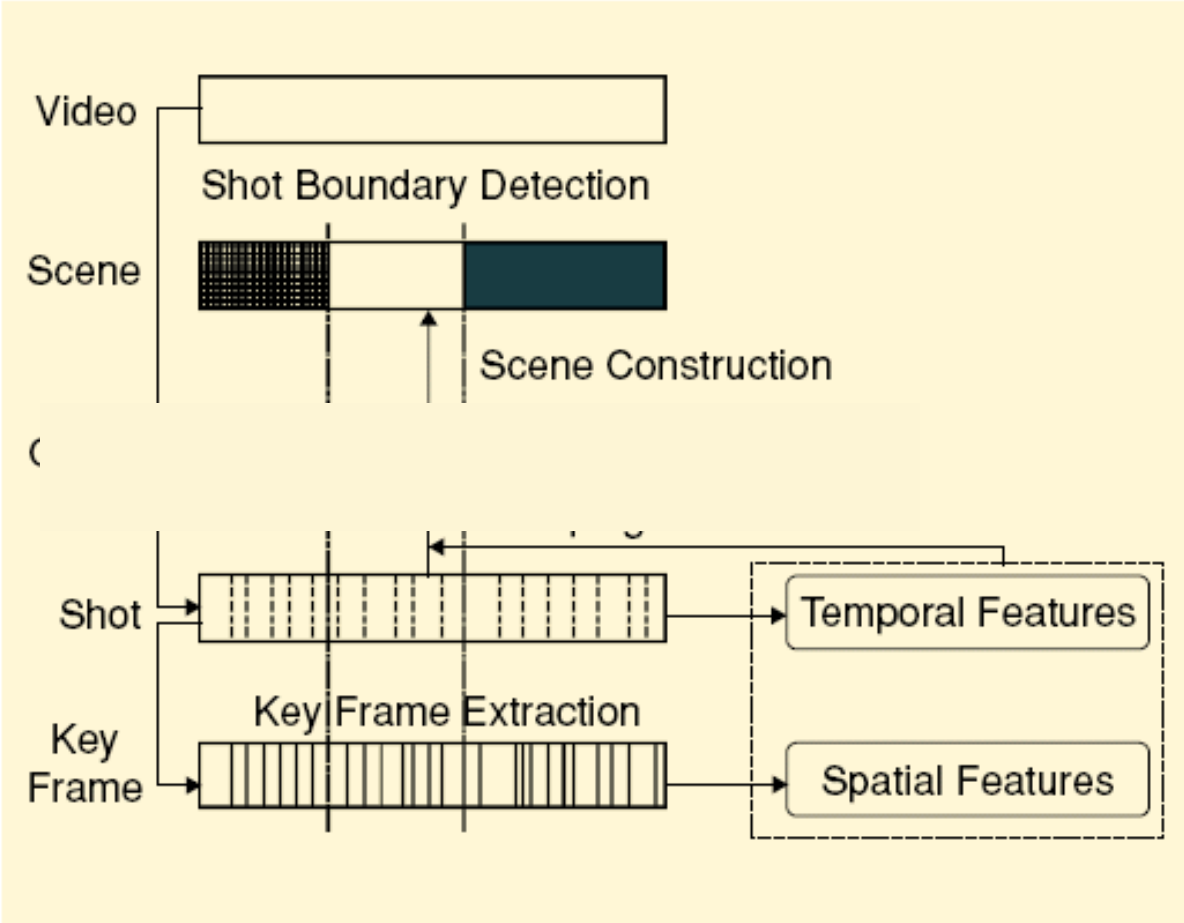
(Images from TRECVID dataset)



Xiong, SPM, 2006

- 视频结构化
  - 结构模型：有脚本视频 vs. 无脚本视频
  - 镜头分割、场景聚类
- 视频高层语义提取
  - 特定对象的检测与跟踪
  - 事件检测
  - 视觉注意建模
- 视频摘要
- 检索模型

- Video
- Scene
- Shot
- key frame



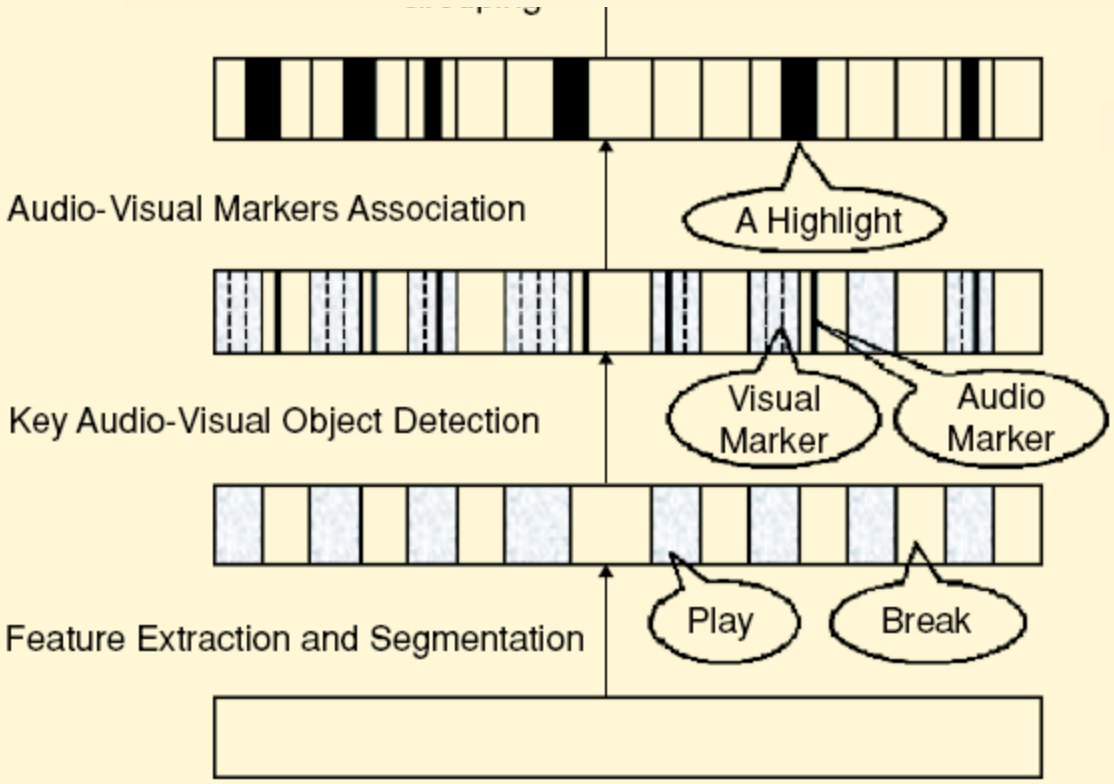
[FIG2] A hierarchical video representation for scripted content.

Xiong, SPM, 2006

play/break

audiovisual markers

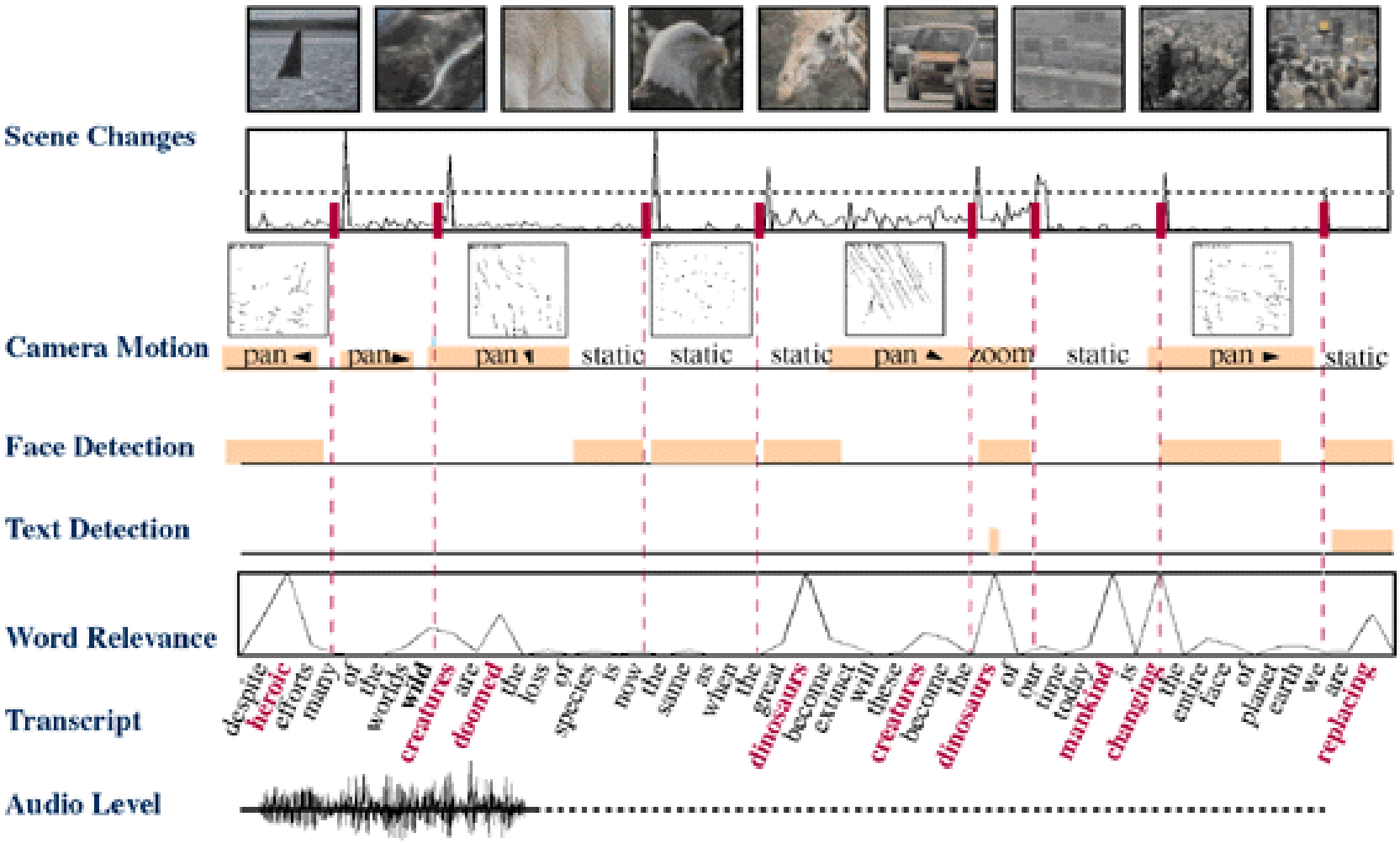
highlights



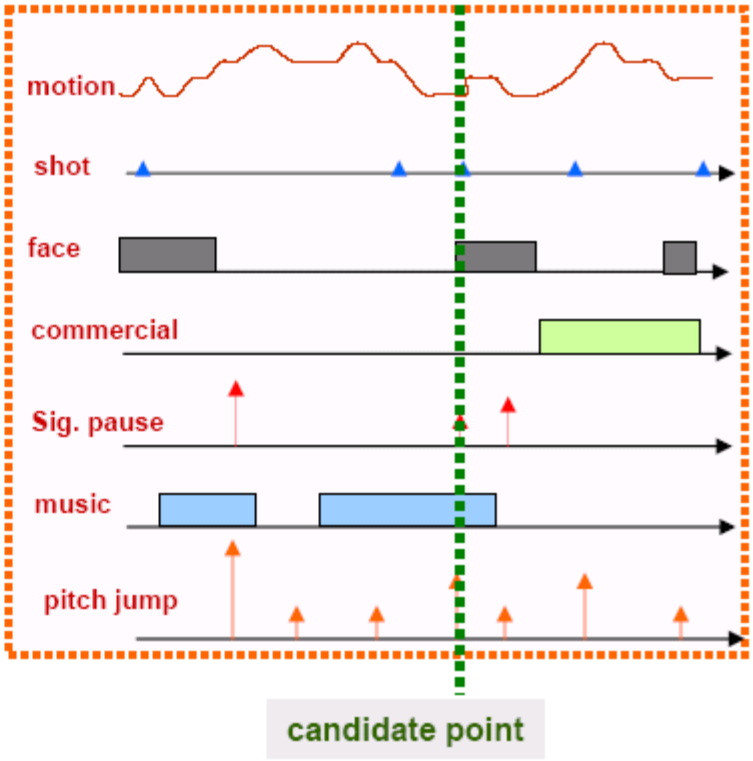
[FIG3] A hierarchical video representation for unscripted content.

Xiong, SPM, 2006

视频检索中可用的特征


Scene Changes
Camera Motion
pan ◄
pan ►
pan ▲
static
static
static
pan ◄
zoom
static
pan ►
static
Face Detection
Text Detection
Word Relevance
Transcript
despite heroic efforts many of the worlds wild creatures are doomed the loss of species is now the same as when the great dinosaurs become extinct will these creatures become the dinosaurs of our time today mankind is changing the entire face of planet earth we are replacing
Audio Level

Modality	Raw Features	Data Type	Value
Video	motion	Point/seg	continuous
	shot boundary	point	binary
	face	segment	continuous
	commercial	segment	binary
Speech /Audio	pause	point	continuous
	pitch jump	point	continuous
	significant pause	point	continuous
	musc./spch. disc.	segment	binary
	spch seg./rapidity	segment	continuous
Text	ASR cue terms	point	binary
	V-OCR cue terms	point	binary
	text seg. score	point	continuous
Misc.	combinatorial	point	binary
	sports	segment	binary



Feature wrappers:

Chang, 2005

Video Caption Extraction in Video Retrieval

Source Video:

Time-Based Minimum Image:

Final VOCR Results:

FREEMAN
BLOCK
LOS
ANGELES
COUNT
SHERIFF

Text Region

Filtered Text

Binarized Segmnted

OCR: S H E R M A N B L O C K

Text Region

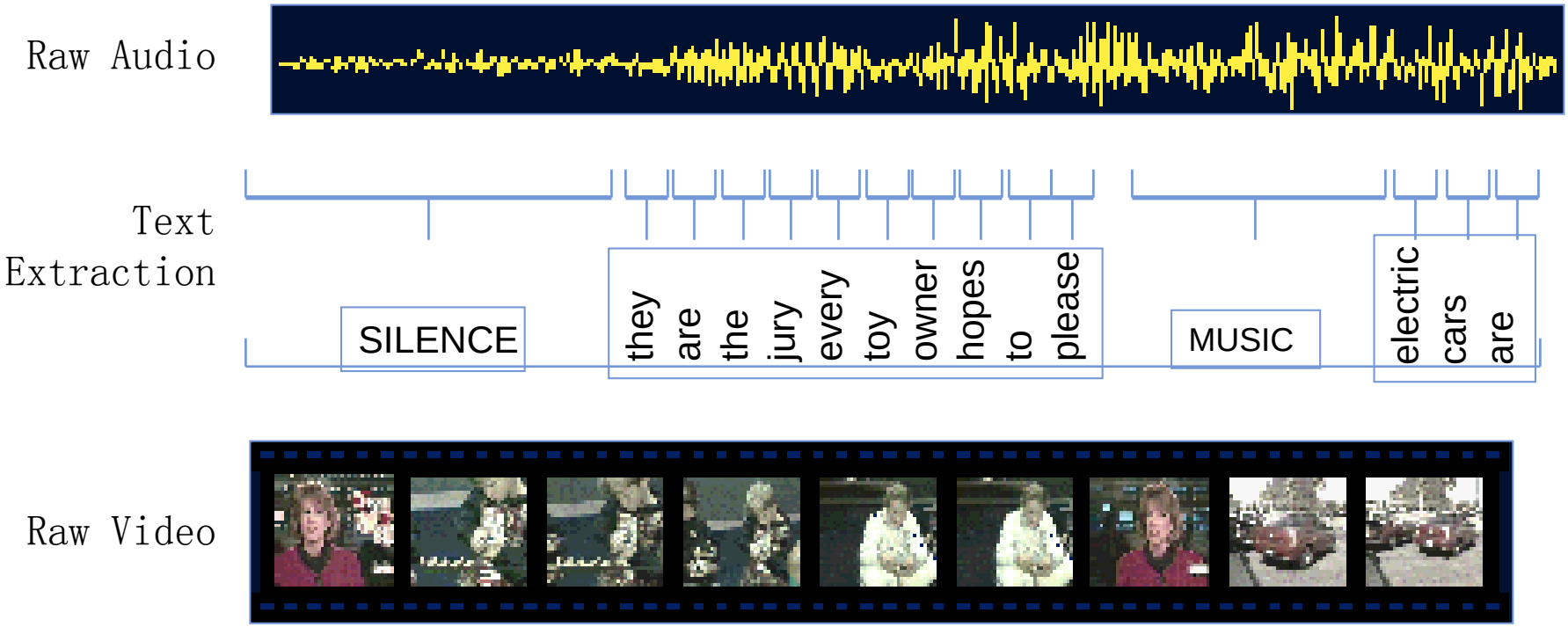
Filtered Text

Binarized Segmnted

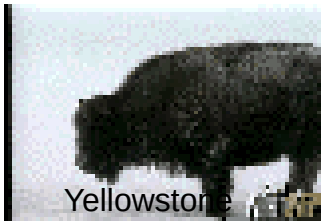
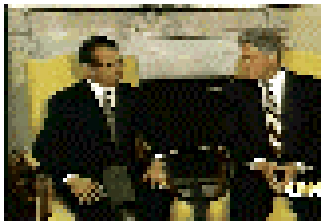
OCR: L O S A N G E L E S C O U N A S H E R I F F

Transcript via Speech Recognition for Video Retrieval

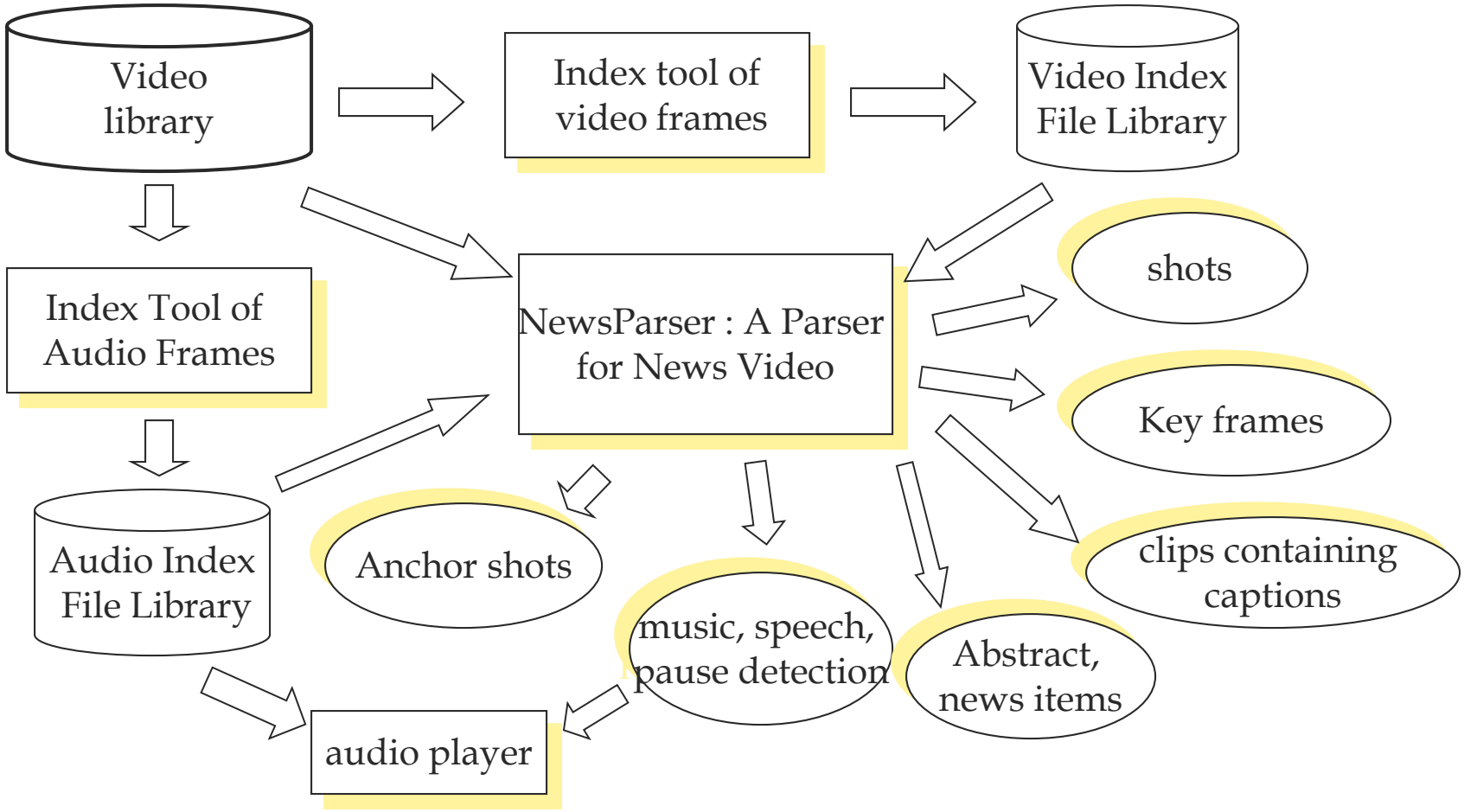
- Generates transcript to enable text-based retrieval from spoken language documents
- Improves text synchronization to audio/video in presence of scripts



Automatic Video Analysis and Index

Scene Cuts			
Camera	Static	Static	Zoom
Objects	Adult Female	Animal	Two adults
Action	Head Motion	Left Motion	None
Captions	[None]	Yellowstone	[None]
Scenery	Indoor	Outdoor	Indoor

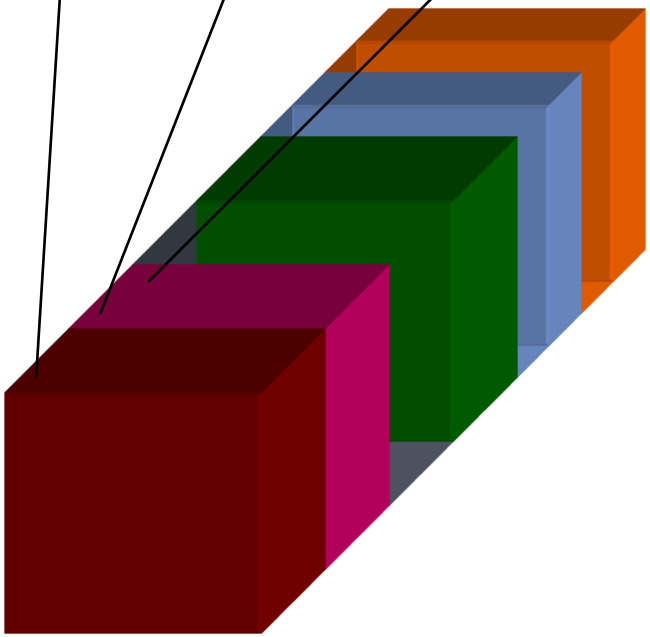
视频(新闻)分析

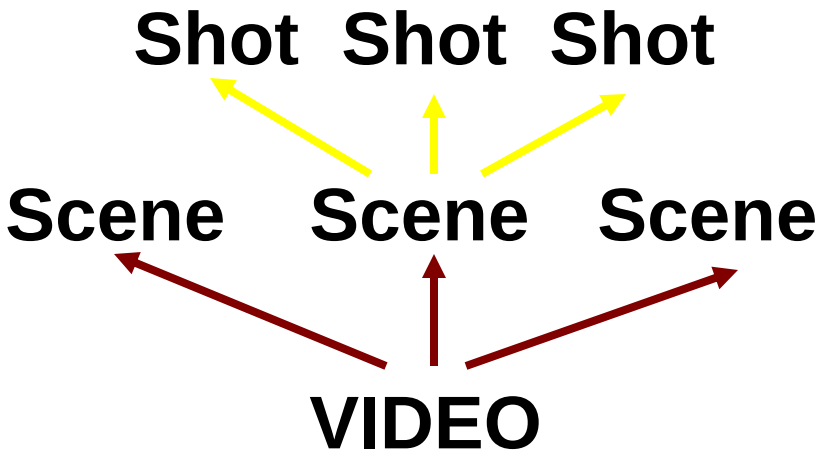

Video library
Index tool of video frames
Video Index File Library
Index Tool of Audio Frames
NewsParser : A Parser for News Video
shots
Key frames
clips containing captions
Audio Index File Library
Anchor shots
music, speech, pause detection
Abstract, news items
audio player

主要困难

- 数据量大
- 无序，非结构化
- 语义信息隐含，内容丰富

Content-based video indexing





内容

- 镜头分割
- 镜头表示
  - 运动目标的半自动分割
  - 背景图象的拼接
- 运动分析
- 视频浏览与检索

视频镜头检测

- 镜头是摄像机在一次连续操作期间拍摄所得的视频帧序列;
- 一个镜头内所有图象描述的应当是比较一致的内容，可以把镜头作为基本索引单元

镜头切换类型

两类镜头切换:

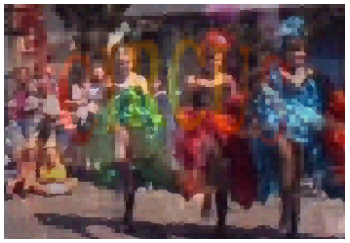
1. 突变：两镜头直接衔接

2. 渐变：两镜头间通过特技操作平滑过渡，包括渐隐，淡入，淡出，扫换等

镜头切换例子




突变




渐变

渐变镜头的一些实例

Dissolve

Wipe

镜头检测的基本原理

- 基本假定：一个镜头内的相邻帧间有较强的连续性和相似性，内容不会有大的变化
- 选择合适的帧间差别测度和合适的阈值，当相邻帧图象间的差别大于阈值时，就认为出现了镜头切换
- 对帧间差别测度的要求：

  对镜头切换敏感

  对镜头内图象的变化不敏感

镜头检测的主要方法

- 基于像素差的方法；
- 基于统计量的方法；
- 基于图象特征的方法；
- 基于灰度或彩色直方图的方法；
- 区域块法；
- 时空流法；
- 压缩域中的方法；
- ……

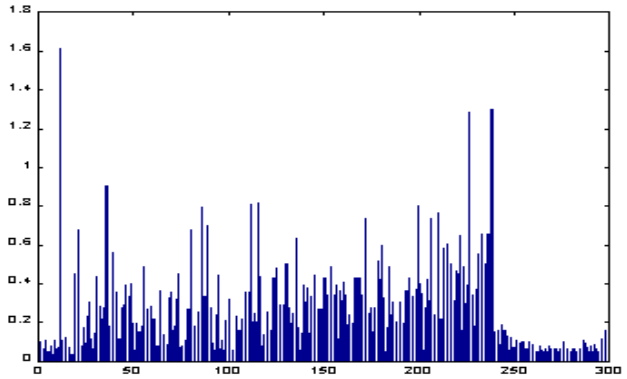
常用测度：颜色直方图

- 镜头内图象变化的原因：运动，光照，小噪声

- 颜色直方图对目标运动和小噪声不敏感，因此得到广泛应用

颜色直方图的缺点

- 颜色直方图对光照变化非常敏感，简单的光强变化就会引起直方图的突变



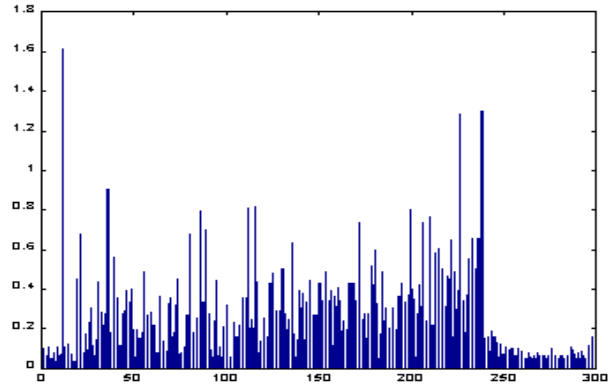
光照变化的例子

一种光照不变测度：颜色比值直方图

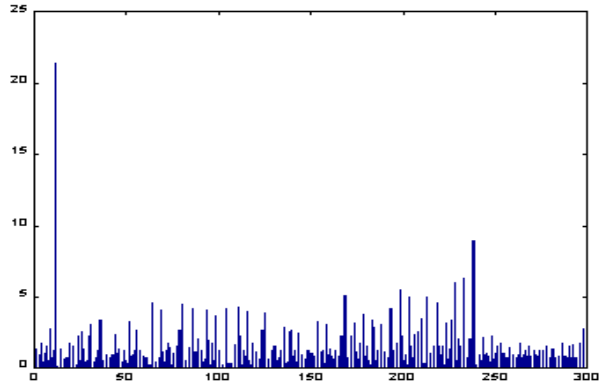
- 相邻象素颜色的比值在光照变化时是不变的
- 颜色比值直方图的差可以作为帧间差别测度

比较

普通颜色直方图



颜色比值直方图



渐变检测

- 渐变检测更困难，因为是平滑过渡，发生切换时相邻帧间仍保持了连续性

- 双阈值技术：低阈值检测可能的起始帧，后续帧与此起始帧比较，高阈值检测渐变结束帧

动态阈值技术

- 镜头切换是视频的一个局部过程，不应采用单一的全局阈值

- 动态自适应的阈值选择

  1. 对当前帧，选择之前的一个时间窗口
  2. 计算这个窗口中帧间差值的均值 a 和方差 σ
  3. 设定双阈值中，低阈值为 $a+(2\sim 3)\sigma$

  $a+(5\sim 6)\sigma$


KF-0001
KF-0002
KF-0003
KF-0004
KF-0005
KF-0006
Vertical Slice (UP) AND Horizontal Slice (DOWN)
Frame Squences
VSLICE
HSLICE
RESET
SAVE
LOAD
UpBound
120
LowBound
10
Scene Description:
"Forest Gump"
Frame
6224
Progress

视频镜头的表示

- 关键帧(Key frame)表示；
  - As Y.T.Zhang, etc. in Proc. ICIP 1998
  - As P.O.Gresles, T.S.Huang, in ICVIS 1997
- 基于图象拼接(Mosaic)的表示
  - M.Irani, P.Anandan and S.Hsu. In ICCV 95
- Highlight表示
  - M.A.Smith, T.Kanade, in CVPR 97

图象拼接方法的优点

- 拼接图加上运动目标包含了镜头的全部内容
- 极大减少了数据量
- 前景与背景分离，容易实现面向对象的操作

拼接图的例子




拼接图的例子（继续）

建立拼接图的方法

- 选定一个参考帧
- 计算其它所有图象与参考帧之间的坐标变换
- 将其它图象变换到参考帧坐标系上，得到拼接图

关键：计算坐标变换，即摄像机运动估计

摄像机运动模型

- 简单平移模型(两参数)
- 平面模型(四参数)
- 仿射模型(六参数)
- 简化透视投影模型(八参数)

简化透视投影模型(八参数)

考虑如下两种情况：

1. 摄像机纯旋转

2. 平面场景

图象间的变换可由8 参数平面射影模型精确描述

$$x' = \frac{a*x+b*y+c}{g*x+h*y+1}$$

$$y' = \frac{d*x+e*y+f}{g*x+h*y+1}$$

模型参数估计

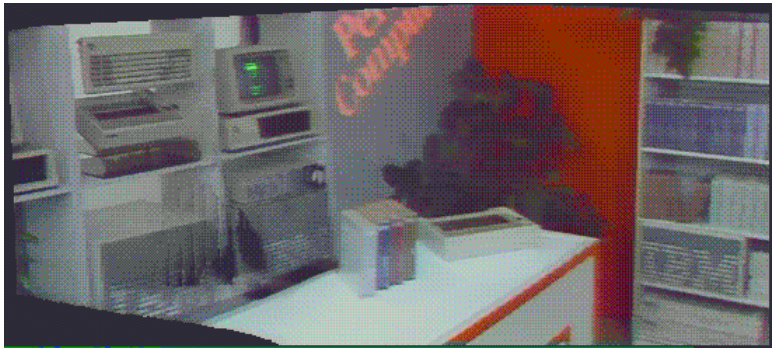
- 基于特征对应的方法
- 无特征对应的直接法

结果演示1

Mosaic Representation

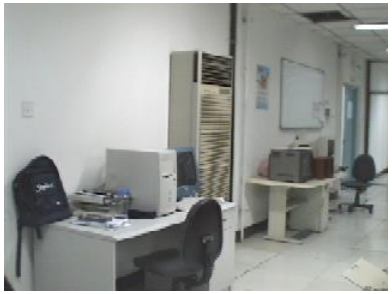
Key Frame Representation

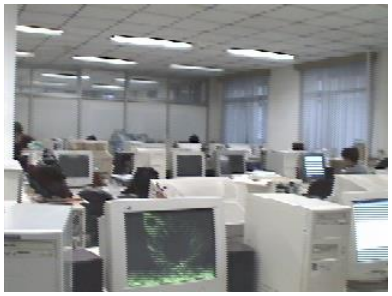
结果演示2



方法一

对于稳定的图像序列，相邻的图像帧之间重叠部分较多，可采用基于Manifold Projection的快速的图像拼接技术

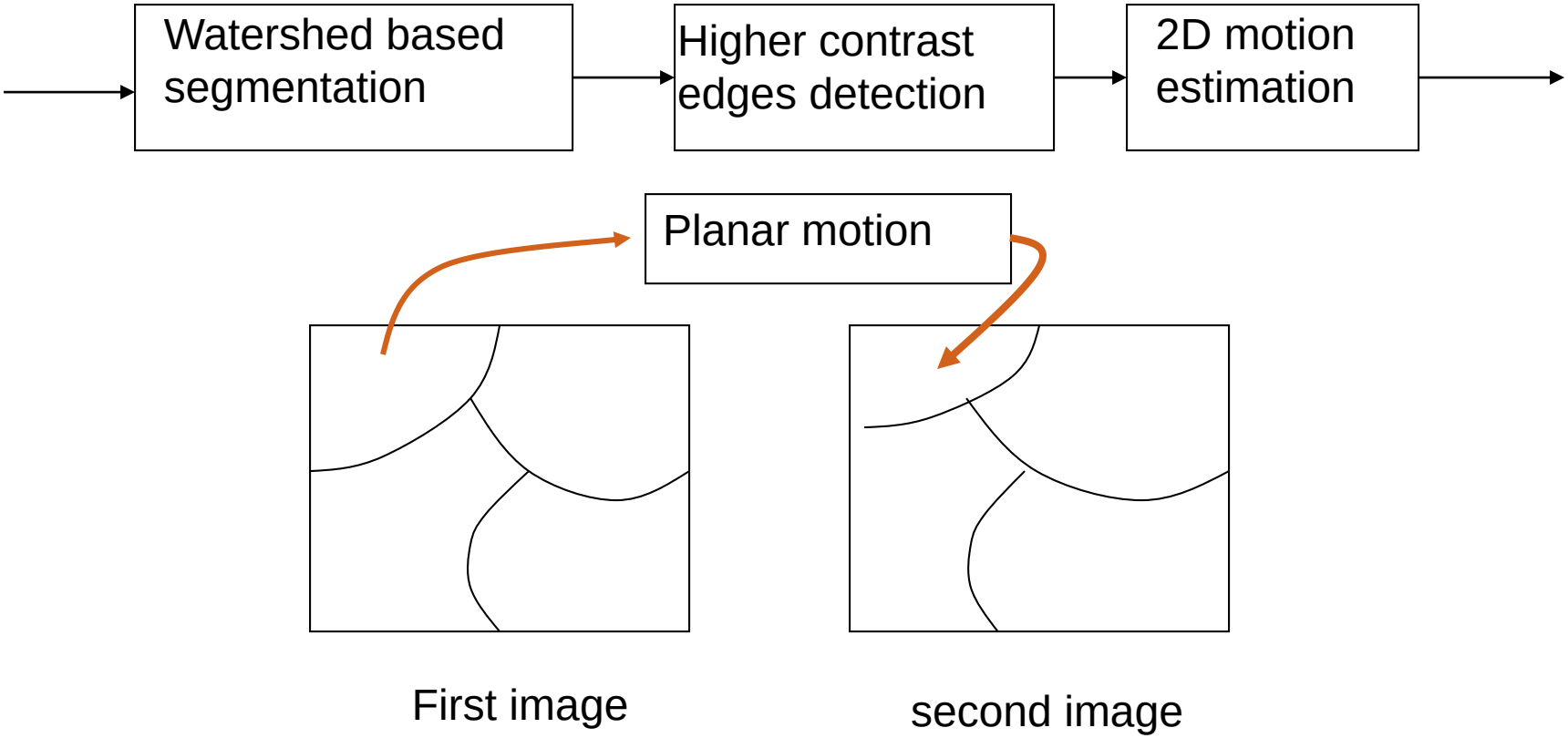




方法二

对于重叠区域较小的图像，可采用基于特征匹配和遗传算法（Genetic Algorithms）的鲁棒的图像拼接技术

方法三 基于边缘的拼接

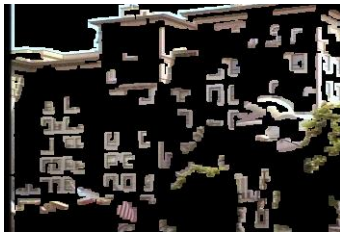

Watershed based segmentation
Higher contrast edges detection
2D motion estimation
Planar motion
First image
second image

算法

- 基于水线的分割
- 高对比度边缘检测
- 鲁棒的平面运动估计
- Mosaic图像生成

高对比度边缘检测

- 高对比度边缘（**Higher contrast edges**）是指具有更多连接边缘的点集
- 膨胀高对比度边缘到给定的宽度



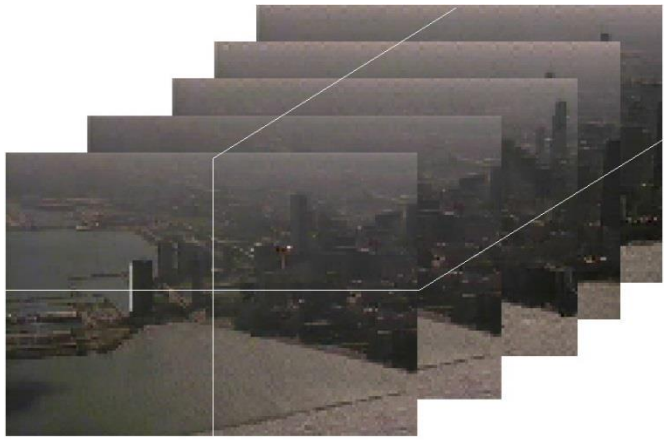
该方法的结果之一

Mosaic Representation

Key Frame Representation

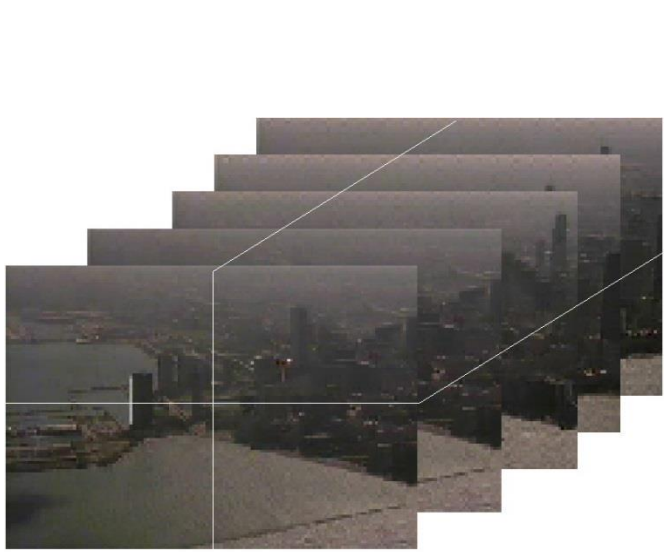
方法四 时空法

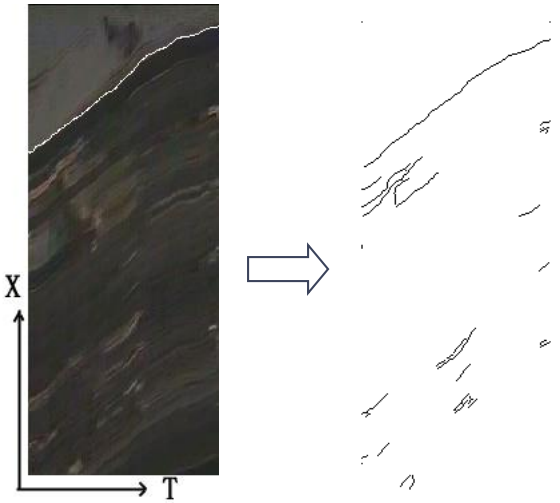
实验一：

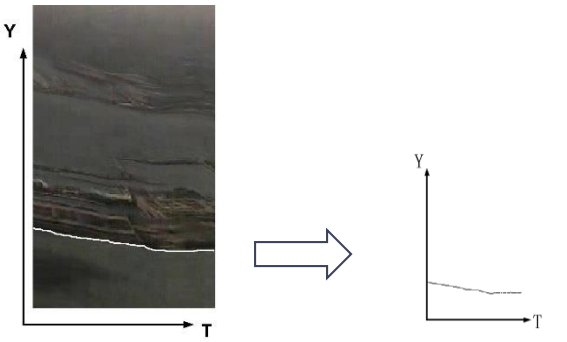


方法四 时空法

实验一：从切片获得的特征曲线







方法四 时空法

实验一：结果比较

用图象金字塔获取初值，最终的结果

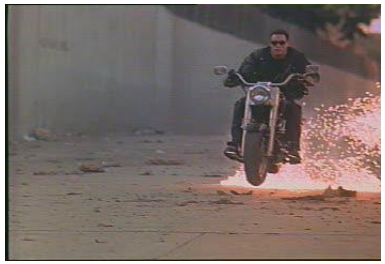
用切片获取初值，最终的结果

方法四 时空法

实验二：存在运动物体的视频段

第20帧

第80帧

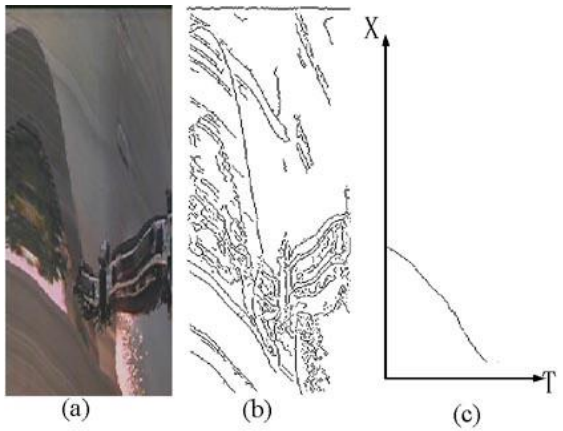


第100帧

第140帧

方法四 时空法

实验二：用我们的方法所得到的结果



运动分析

- 运动物体分割
- 运动物体的识别与跟踪
- 行为动作理解

运动目标分割

- 时域差分法
  - 优点：方法简单快速
  - 缺点：摄像机运动时需要在分割前完成运动补偿
- 运动估计与分割
  - 外在方法
  - 隐式方法

光流场估计

- 基于点亮度不变性
- 基于特征不变
- 基于区域相似性

运动目标的识别与跟踪

- 识别
  - 目标获取、定位、识别.....
- 跟踪
  - 运动参数估计、预测，运动路径描述.....

行为动作理解

- 哑语
- 虚拟鼠标、虚拟环境.....
- 行为动作的意义描述
- .....

应用背景和目标

- 面向基于内容的视频数据编码与索引：MPEG-4和MPEG-7。
- 视频目标提取：根据一定的准则，把视频图象分割成不同的区域并标识出有语义意义的目标。

主要难点

1、对区域的语义描述。
2、摄像机（全局）运动补偿。
3、多运动估计和分割。

策略

- 1、区域的语义描述
- 初始帧的交互式标定加后续帧的自动目标跟踪。
- 2、摄像机（全局）运动补偿
- 与视频目标提取相结合，利用语义信息。
- 基于多尺度梯度水线的特征匹配方法。
- 3、多运动估计和分割。
- 采用了自上而下的空域分裂策略，根据不同的运动复杂度使用不同尺度的空域分割，组成一个区域金字塔，从而提出了一个新的时空域多运动估计和分割方法。

形态学空域分割的目的

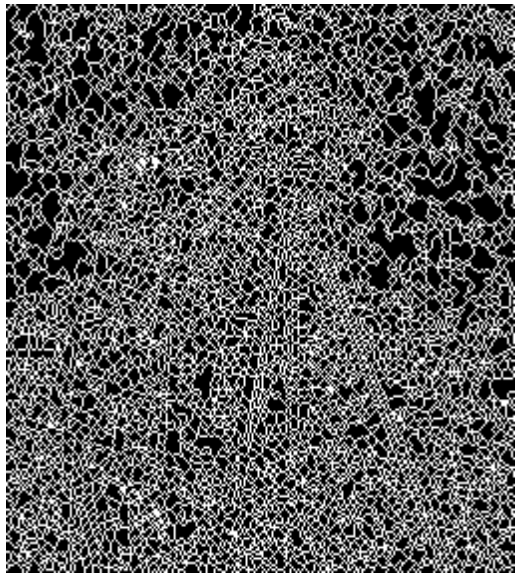
- 给出按某些准则划分出一致性区域，再应用基于区域的编码方法以提高编码效率。
- 为视频目标分割提供一个很好的划分基础和空间拓扑约束。

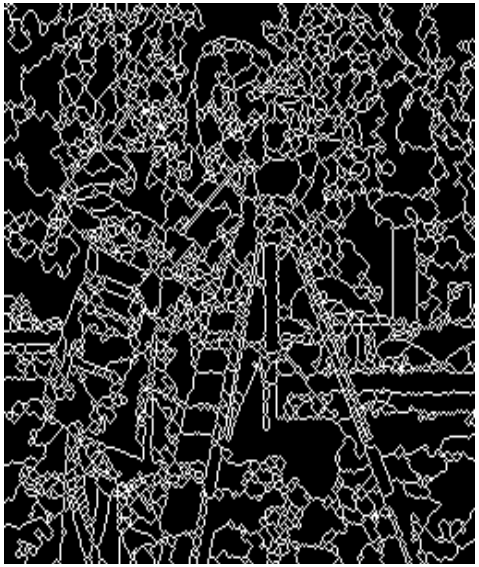
基于数学形态学的层次化图象分割

- 多尺度重建滤波器：在不同尺度下简化图象以利于分割。

- 水线算法给出区域划分。

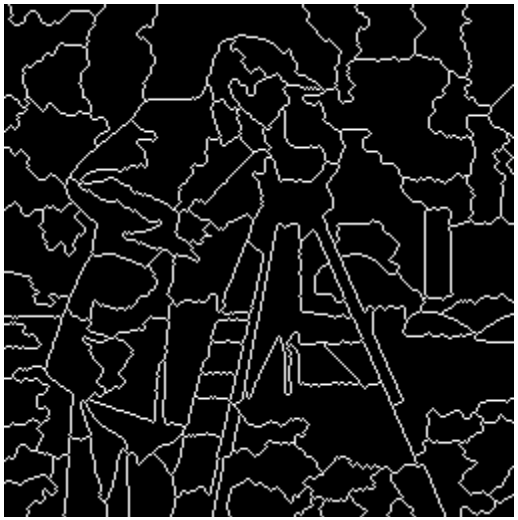
水线分割的过分问题





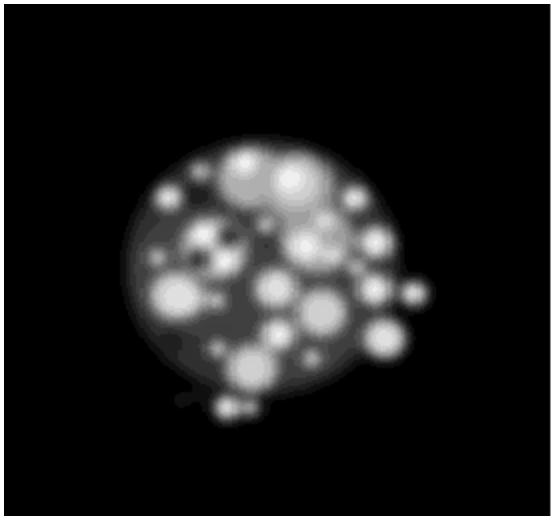
解决方法

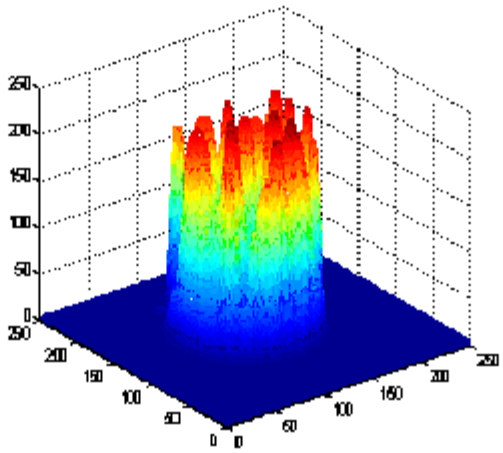
- 利用重建算子修改梯度图象

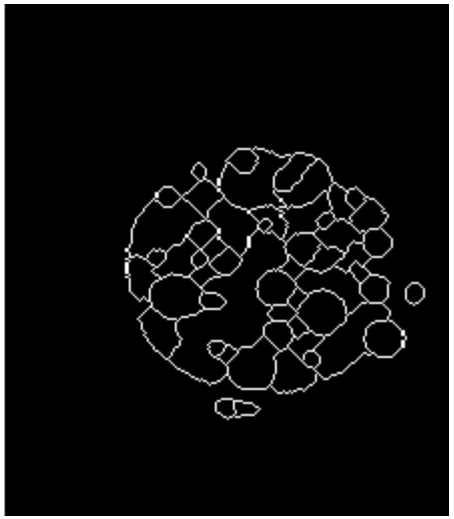


多尺度重建滤波器下的梯度水线

- 具有良好结构对应性的层次化区域划分方法

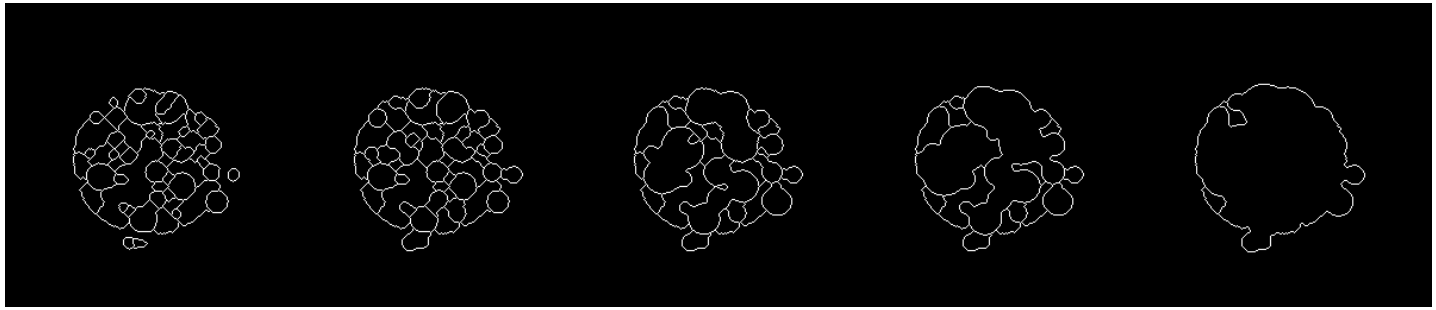




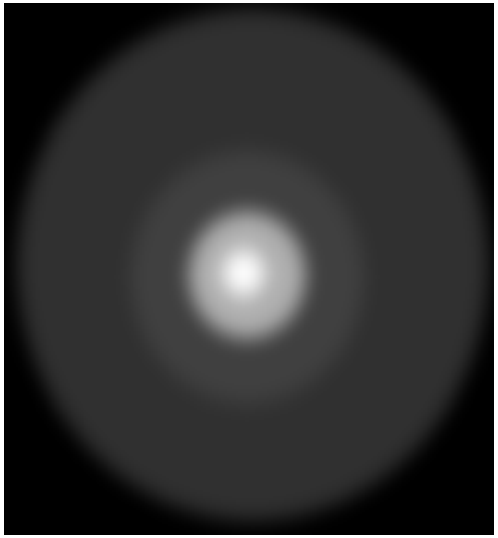


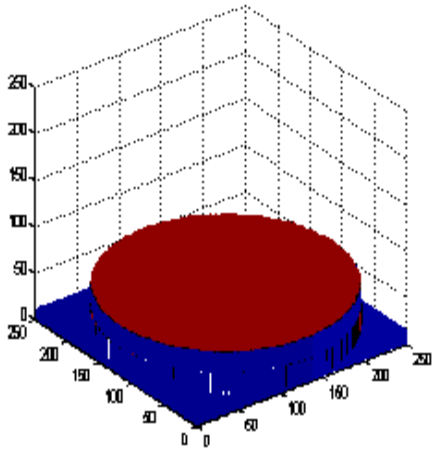
图象的区域金字塔表示

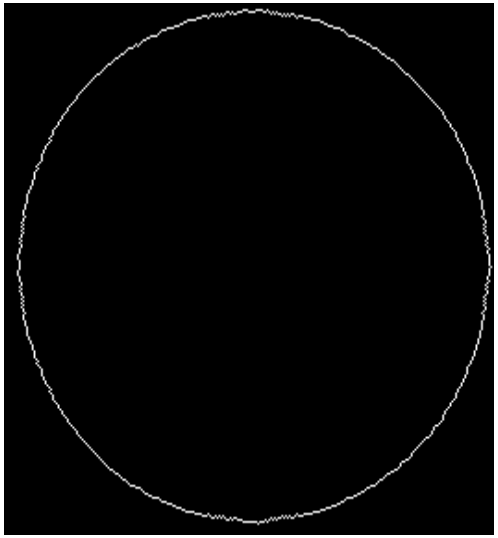
- 二维图象结构特征的尺度化提取方法



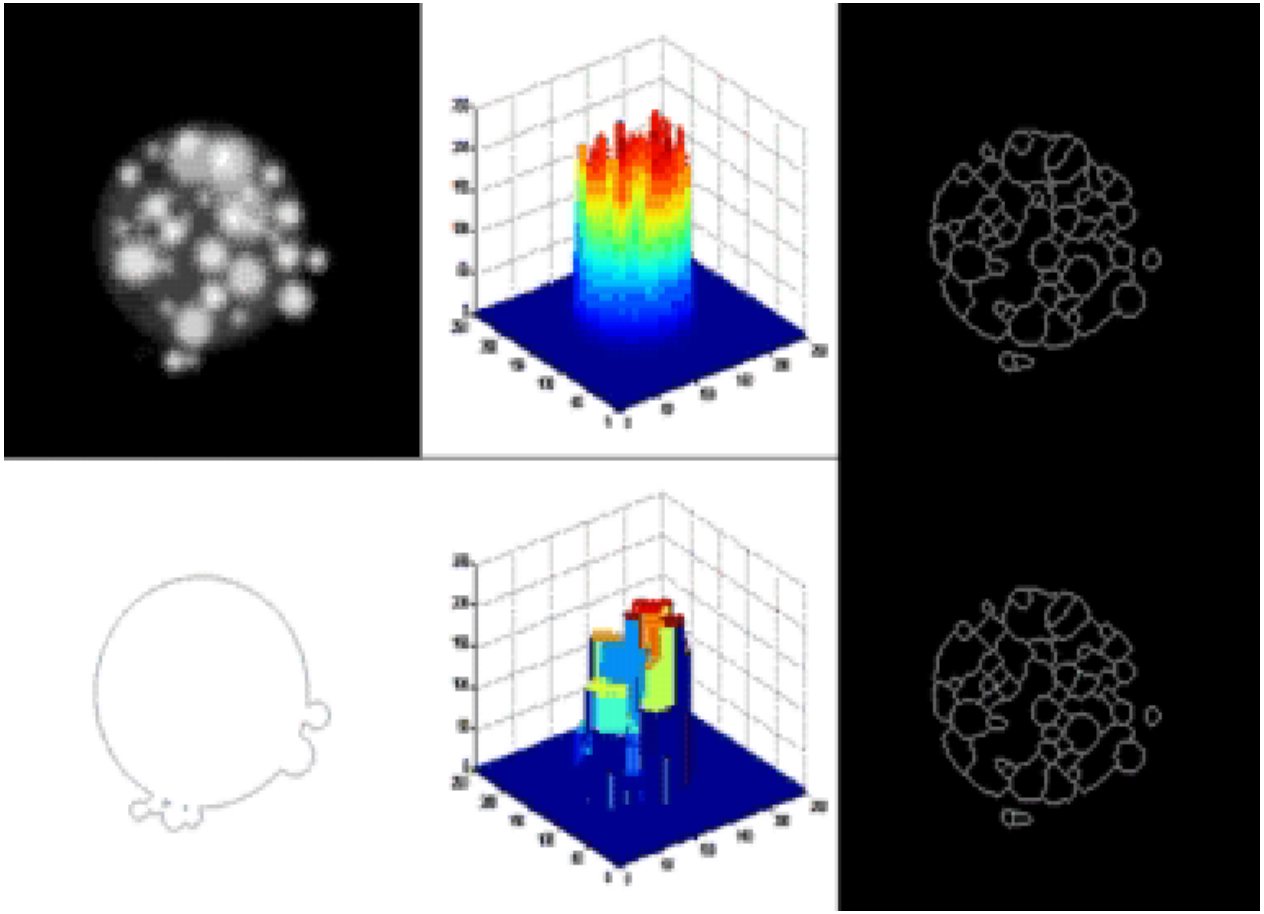
漂移和层化变换



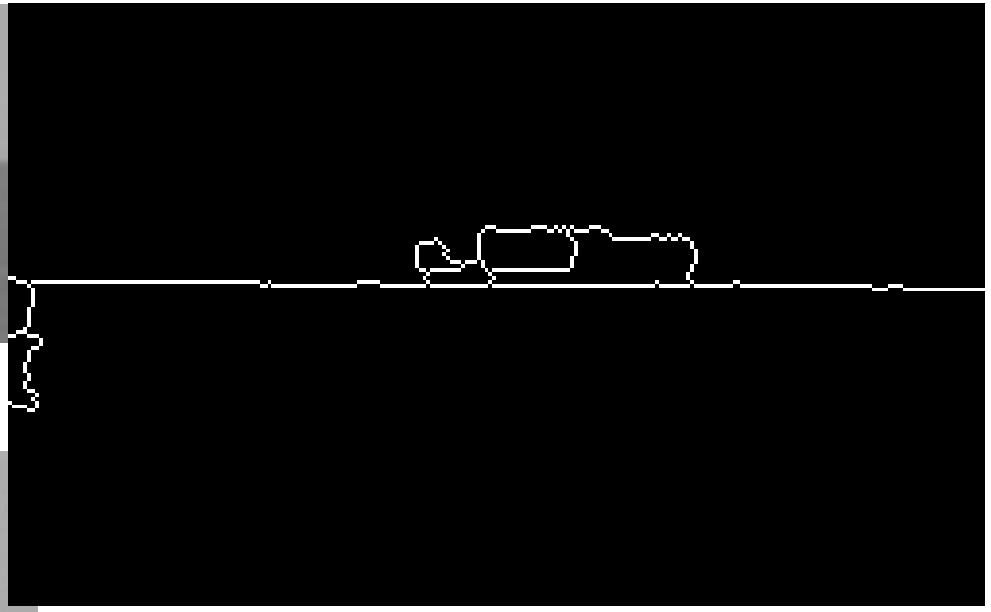


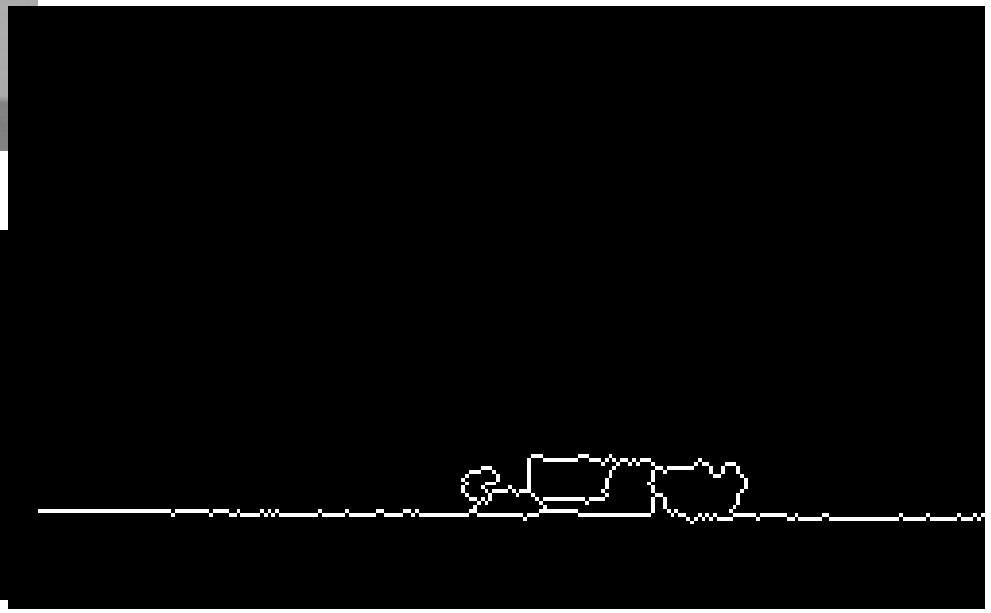


对比

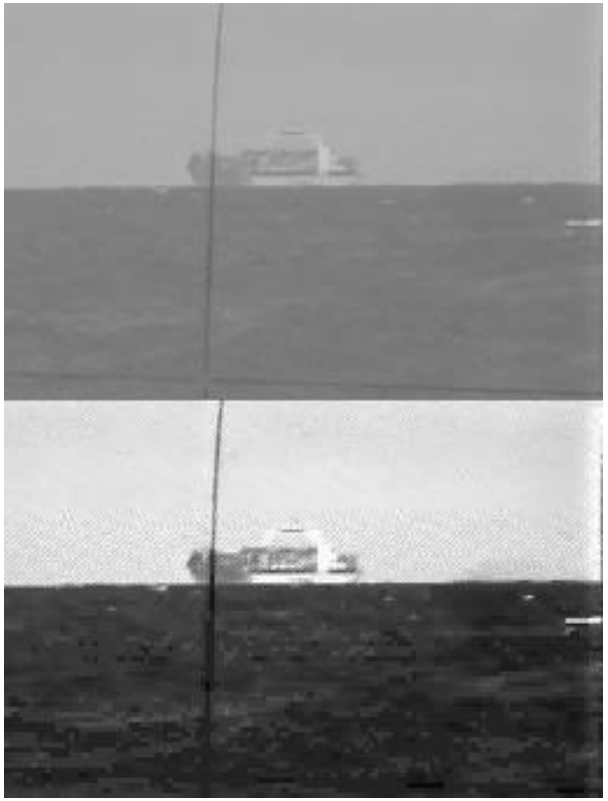


多尺度匹配





实验



多运动分割

- 基本假设：参数化区域运动在拓扑分布上的一致性
- 困难：运动分析的综合问题
- 方法：鲁棒性运动参数估计和运动预测

  基于图象层次化区域表示和一致性运动约束的区域分裂

  基于运动测度的连通形态滤波的区域合并

  区域补偿解决遮挡问题

半自动语义运动分割

- 语义分割:新一代视频压缩的关键技术
- 困难：语义定义的困难和目标形体知识的缺乏，以及多运动分割
- 方法：交互式标定解决初始语义分割

  多尺度形态结构匹配解决全局运动

  基于层次化区域表示的参数化多运动估计解决随后的自动目标跟踪

框架

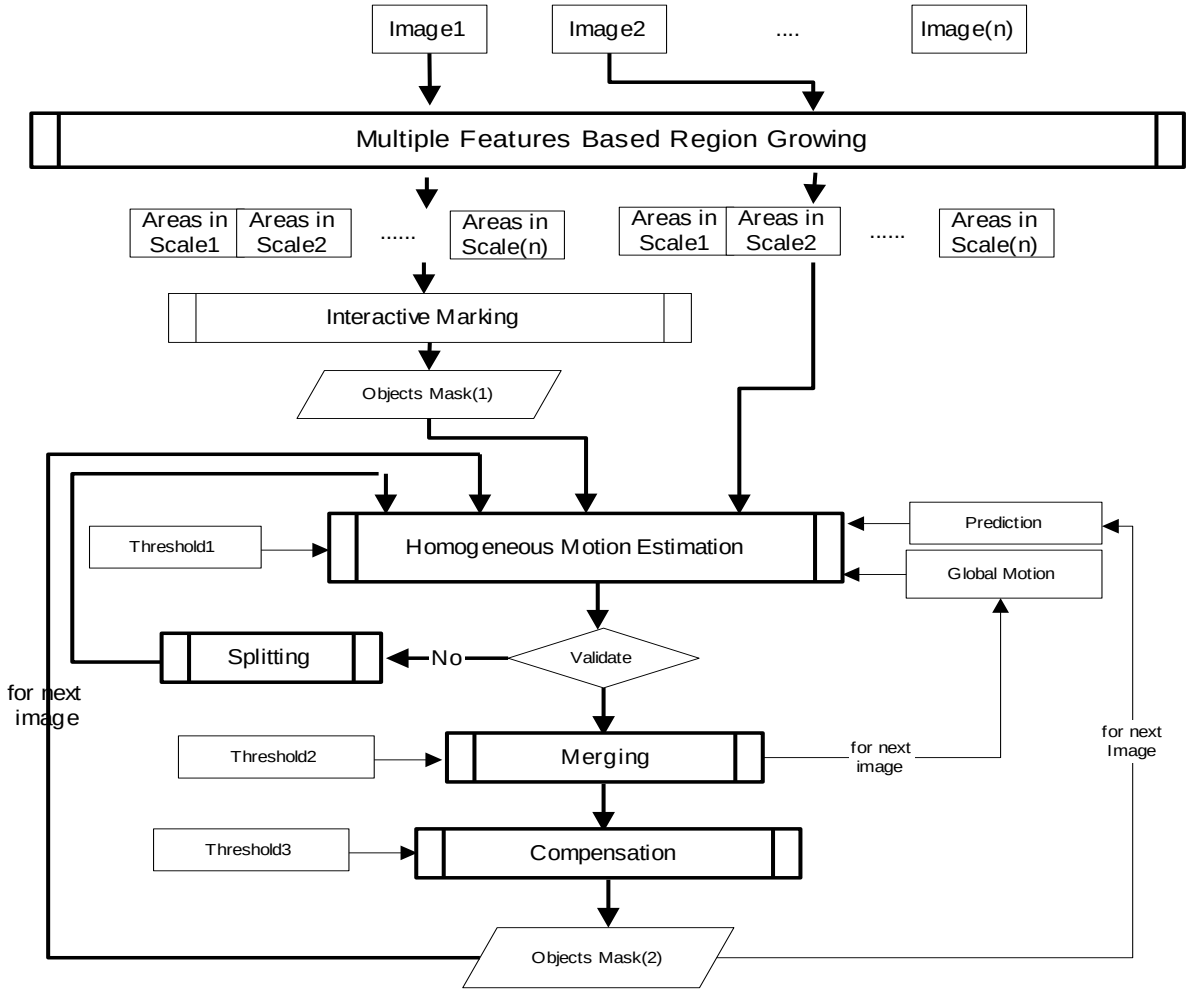
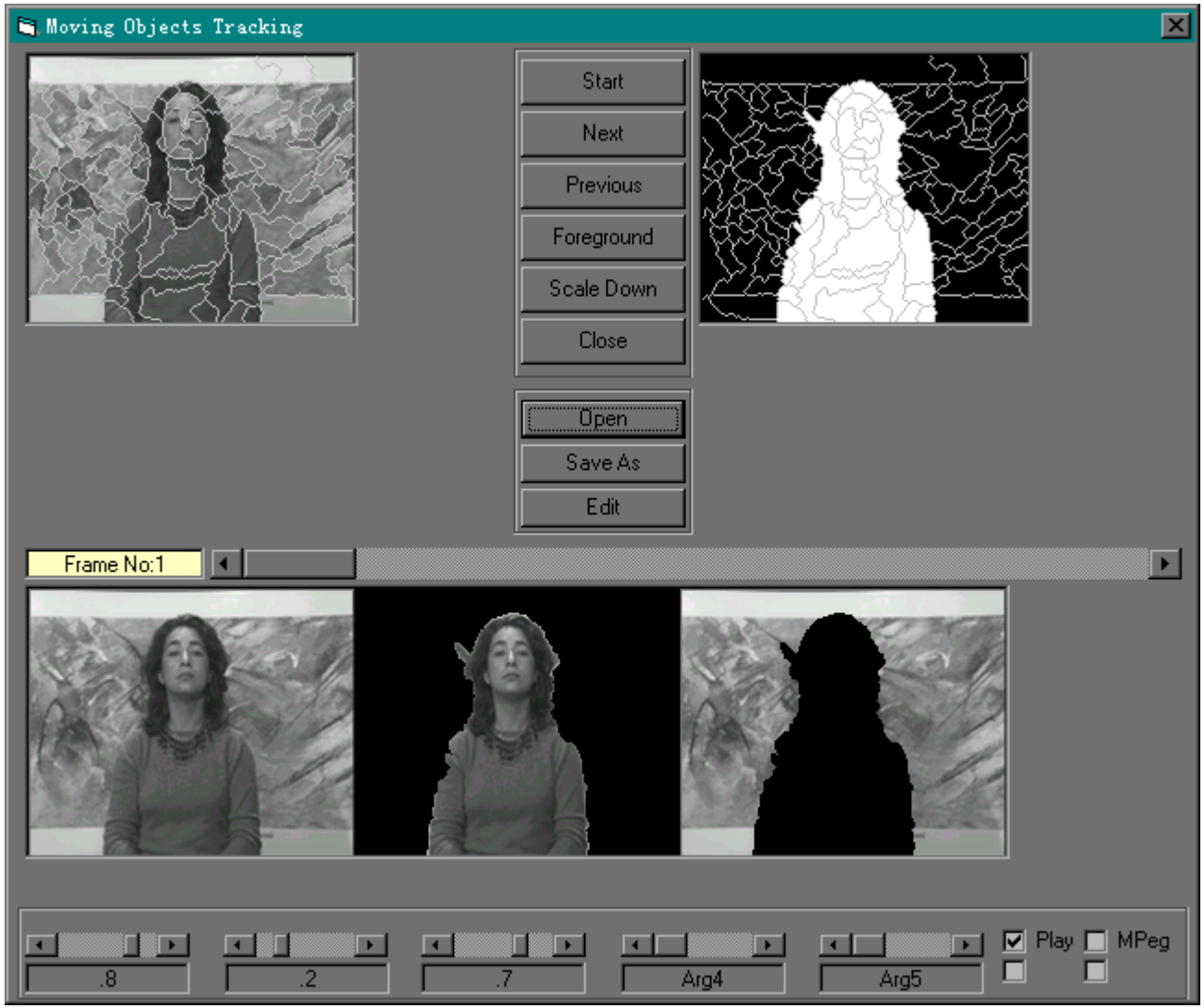
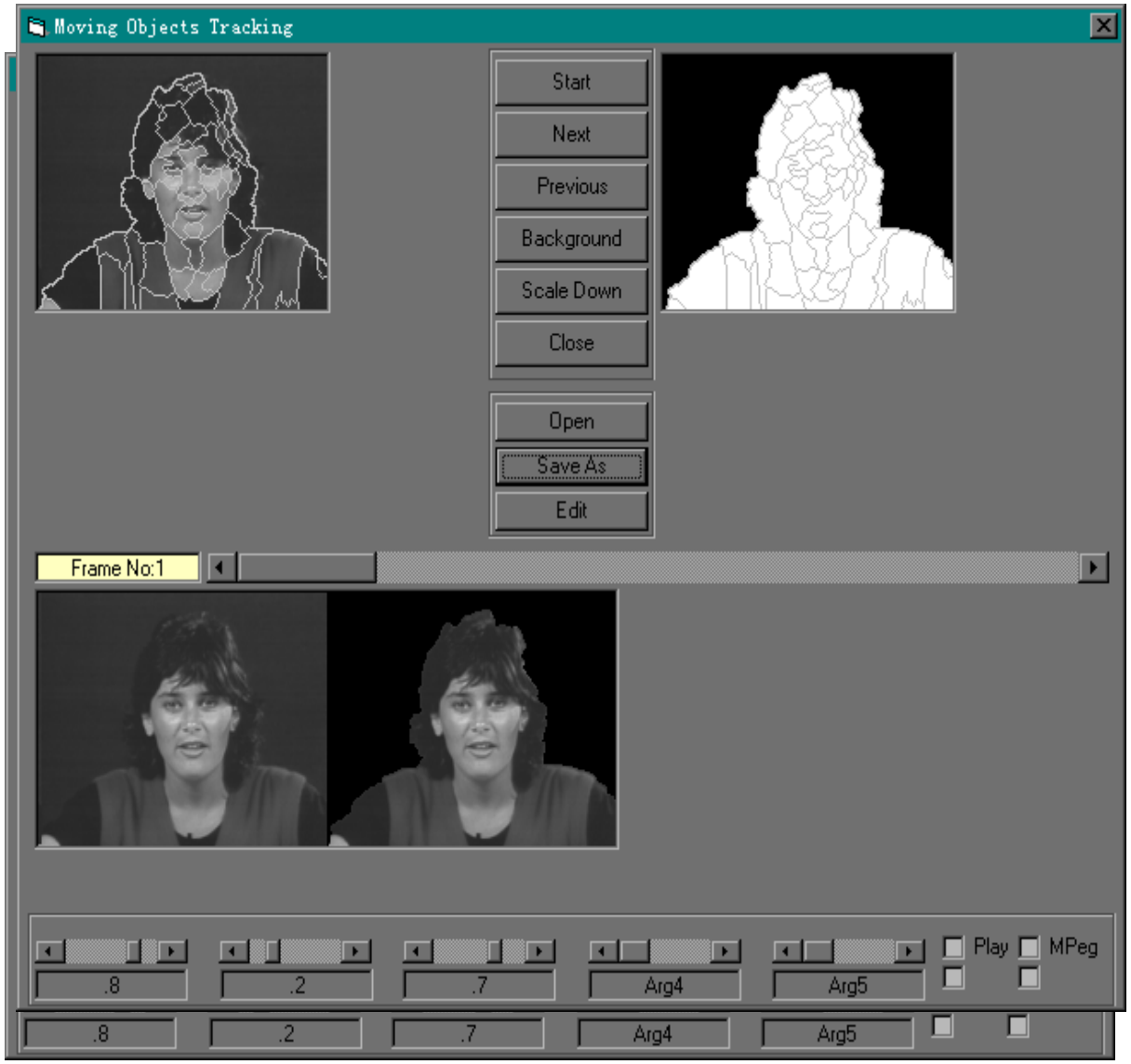


Image1
Image2
....
Image(n)
Multiple Features Based Region Growing
Areas in Scale1
Areas in Scale2
.....
Areas in Scale(n)
Areas in Scale1
Areas in Scale2
.....
Areas in Scale(n)
Interactive Marking
Objects Mask(1)
Threshold1
Homogeneous Motion Estimation
Prediction
Global Motion
Splitting
No
Validate
for next image
Threshold2
Merging
for next image
for next Image
Threshold3
Compensation
Objects Mask(2)

程序界面


Moving Objects Tracking
Start
Next
Previous
Foreground
Scale Down
Close
Open
Save As
Edit
Frame No:1
.8
.2
.7
Arg4
Arg5
Play
MPeg

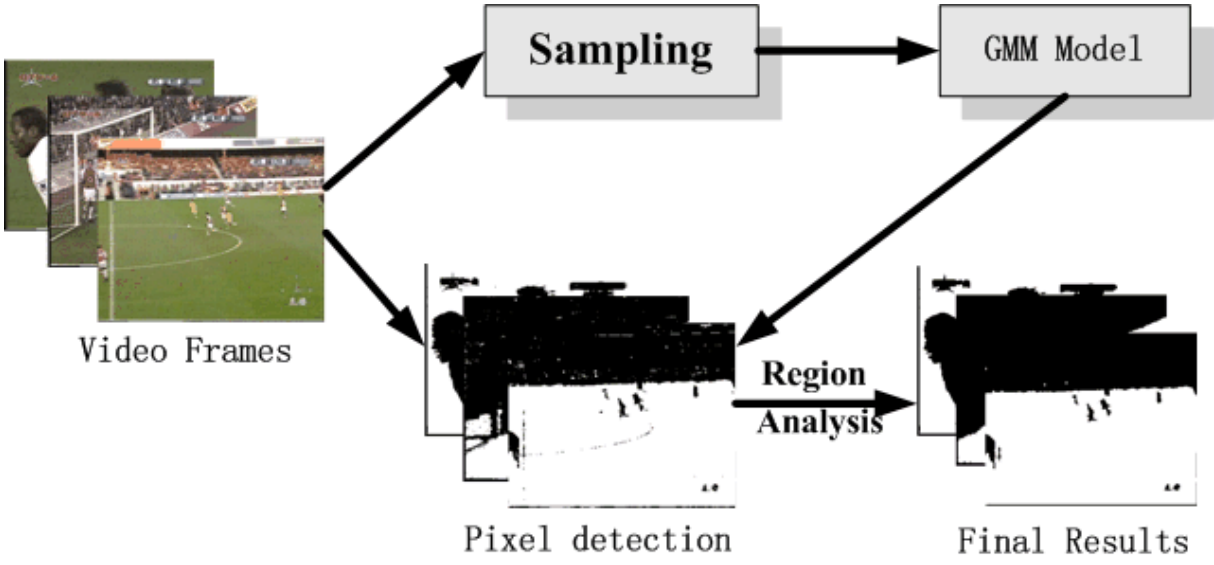
多尺度交互标定


Moving Objects Tracking
Start
Next
Previous
Background
Scale Down
Close
Open
Save As
Edit
Frame No:1
.8
.2
.7
Arg4
Arg5
Play
MPeg
.8
.2
.7
Arg4
Arg5

分割实验1

分割实验2

- Plays the fundamental role in analyzing many kinds of sports videos such as soccer, tennis, basketball, etc
- Playfield often occupies dominant color region and becomes the major part in video frames



- Examples on various playfield detection

(a)

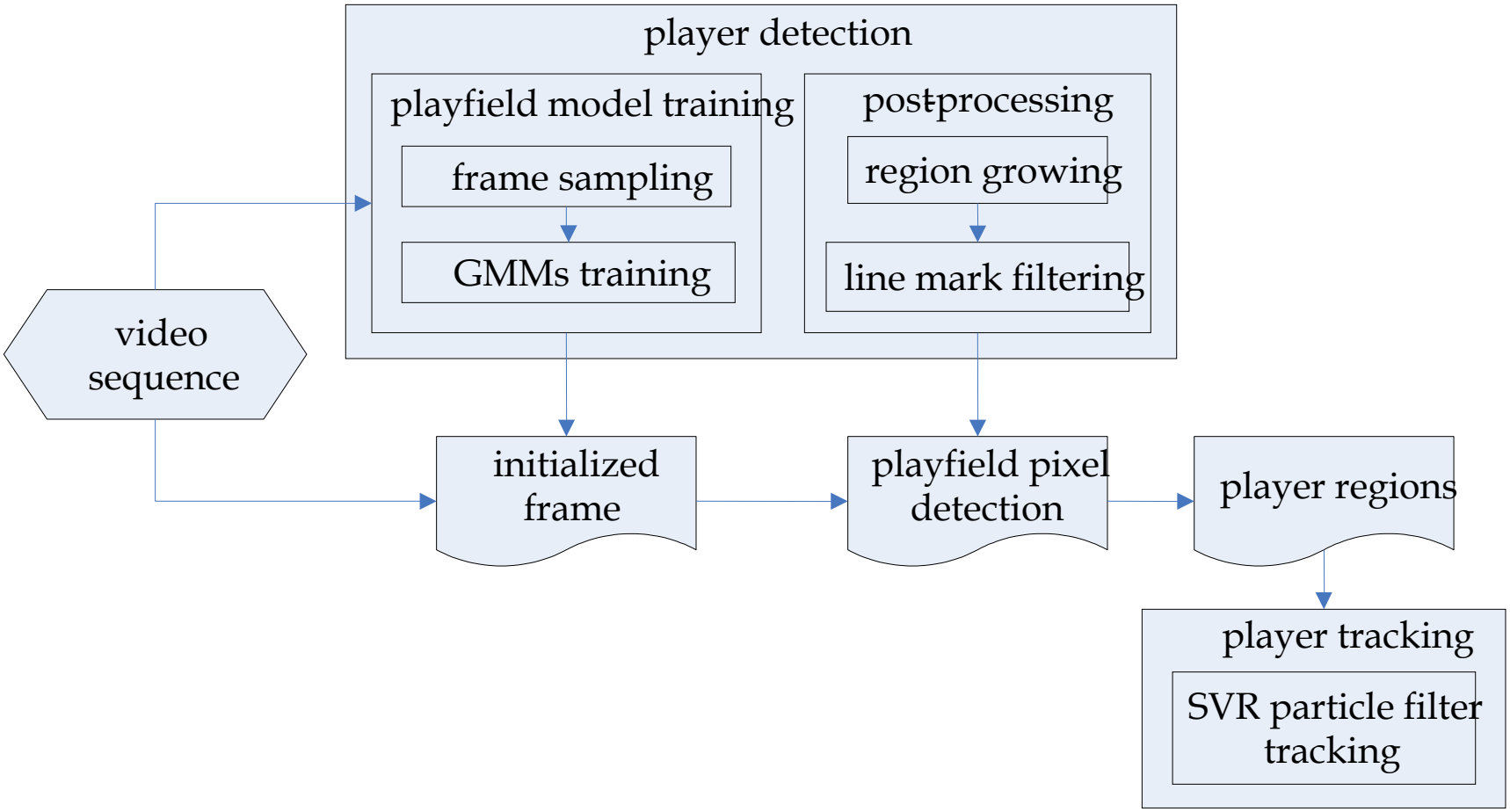
(b)

(c)

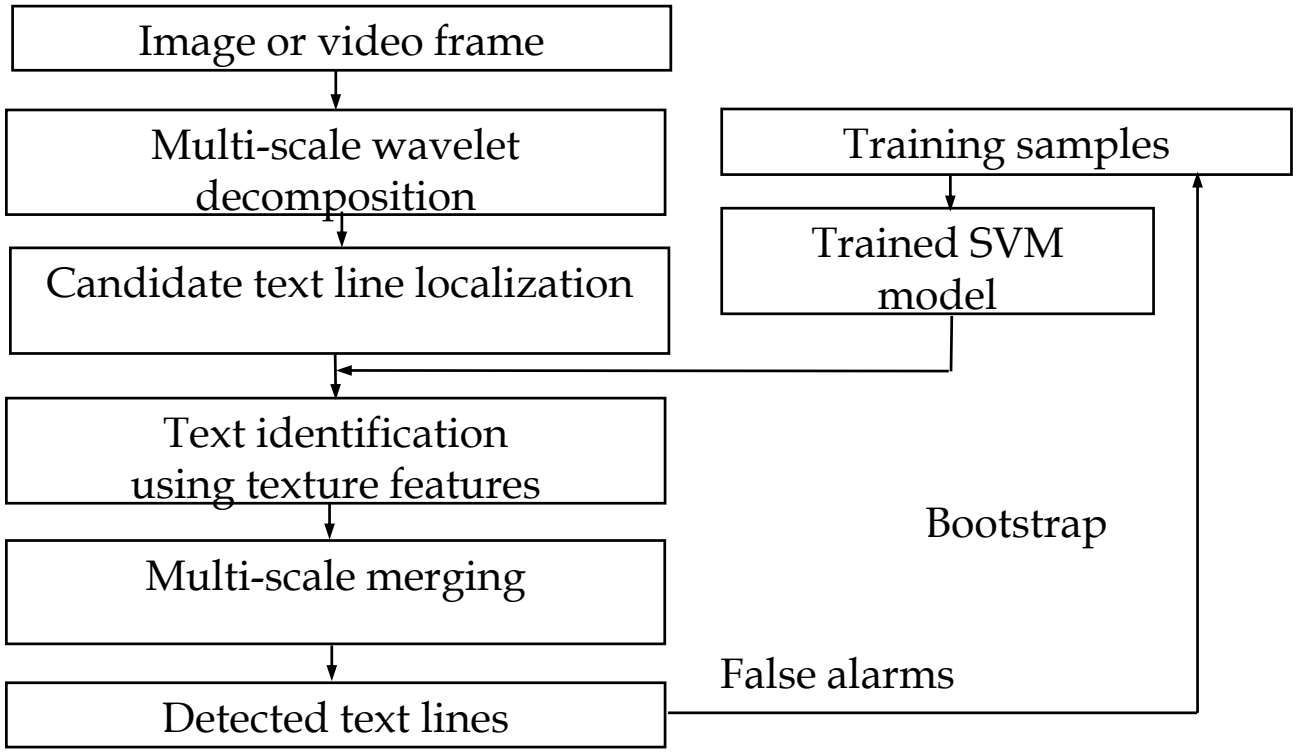
(d)

- Player detection is based on results of playfield segmentation and region analysis
- SVR (Support Vector Regression) particle filter is used for player tracking
  - Solve re-sampling problem in particle filter by SVR-based sample re-weighting

对象检测与跟踪：球员

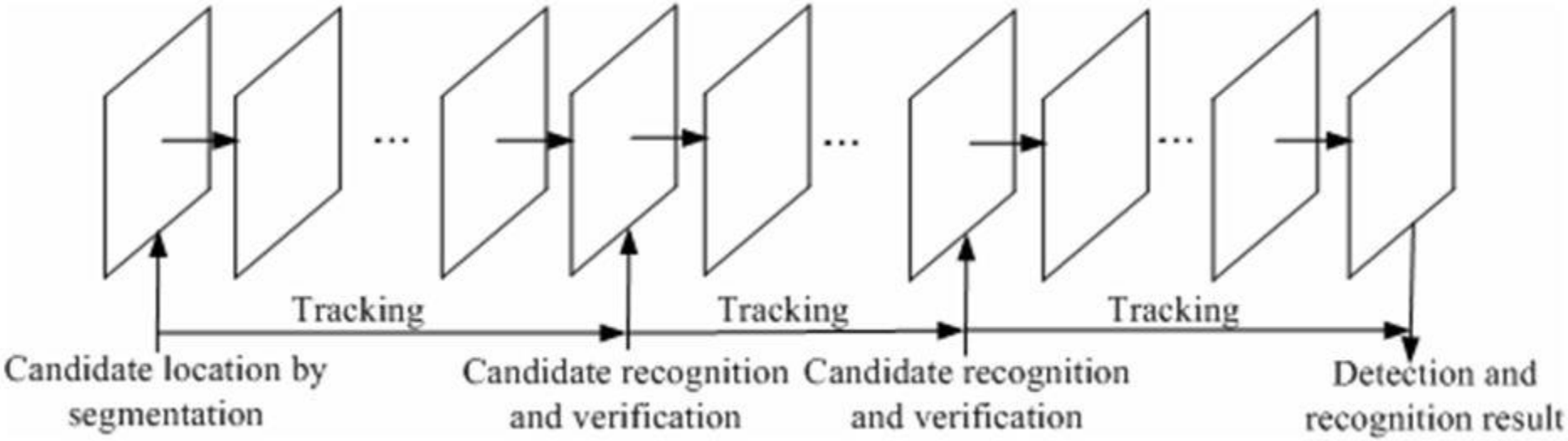

player detection
playfield model training
frame sampling
GMMs training
postprocessing
region growing
line mark filtering
video sequence
initialized frame
playfield pixel detection
player regions
player tracking
SVR particle filter tracking

- **Overlay text contains much semantic information in video**
- **By using multi-scale wavelet features, a novel coarse-to-fine algorithm is proposed to locate text lines even under complex background.**



对象检测与跟踪：文字

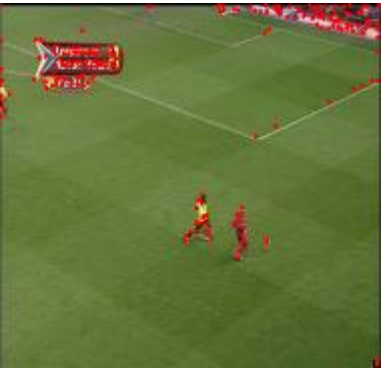
对象检测与跟踪：文字



* Published in VCIP2005

视频结构化：镜头边界检测

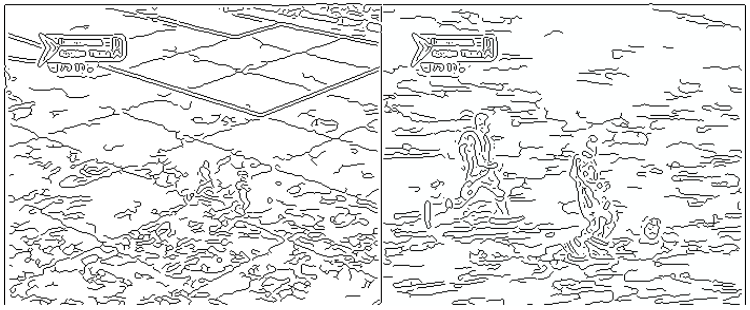
- Illustration



328

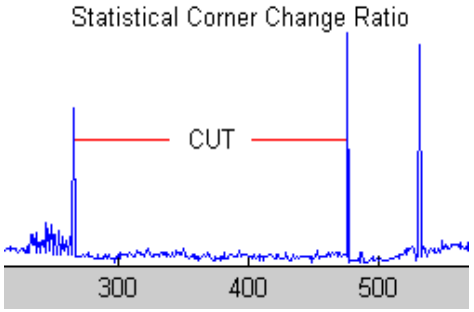
329

Harris corner detect results

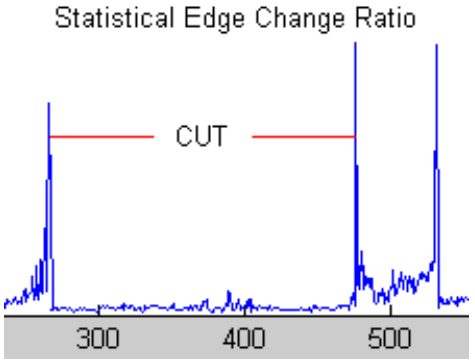


328 329

Canny edge detect results



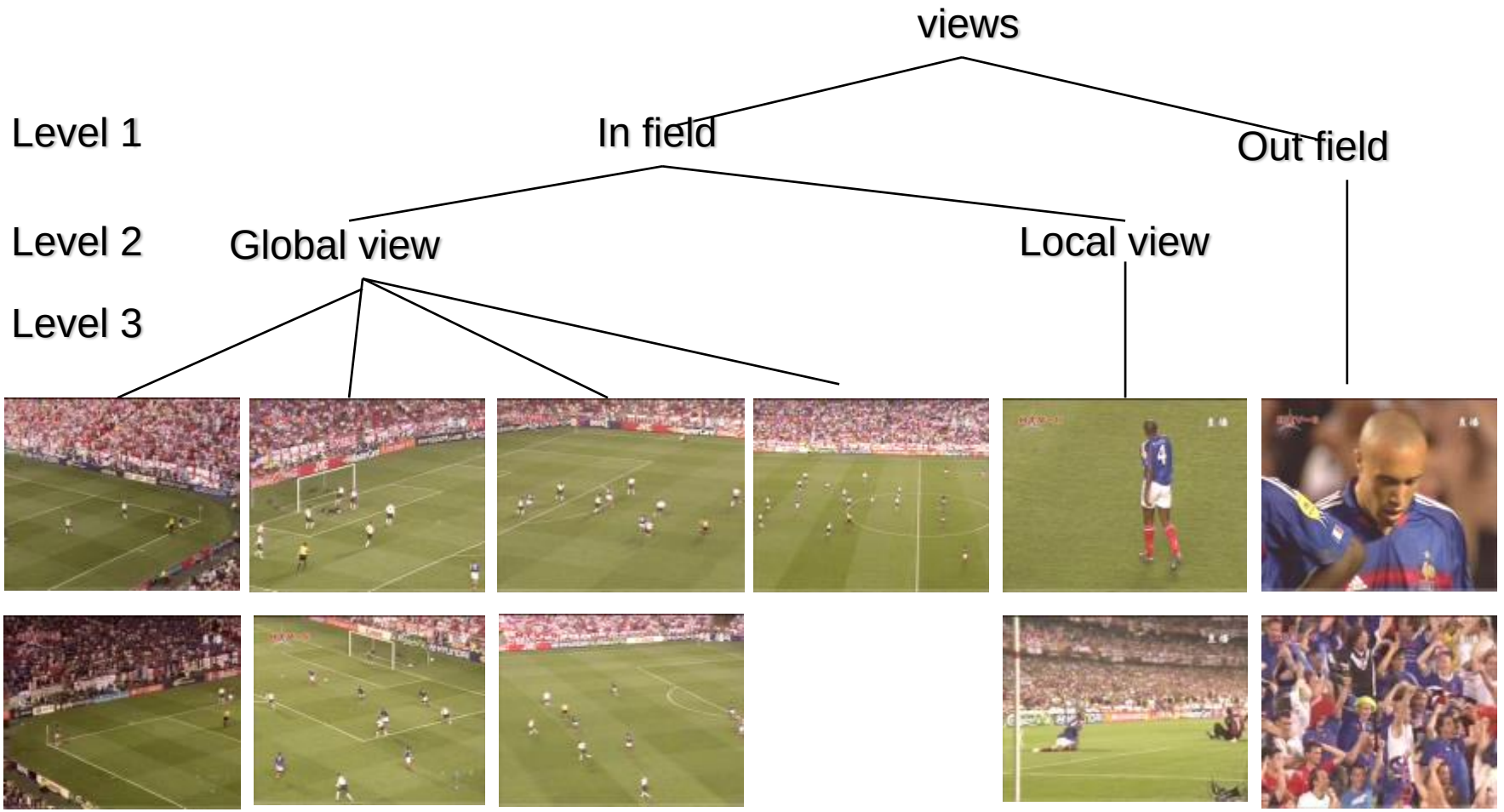
shot detection result using SCCR



shot detection result using SECR

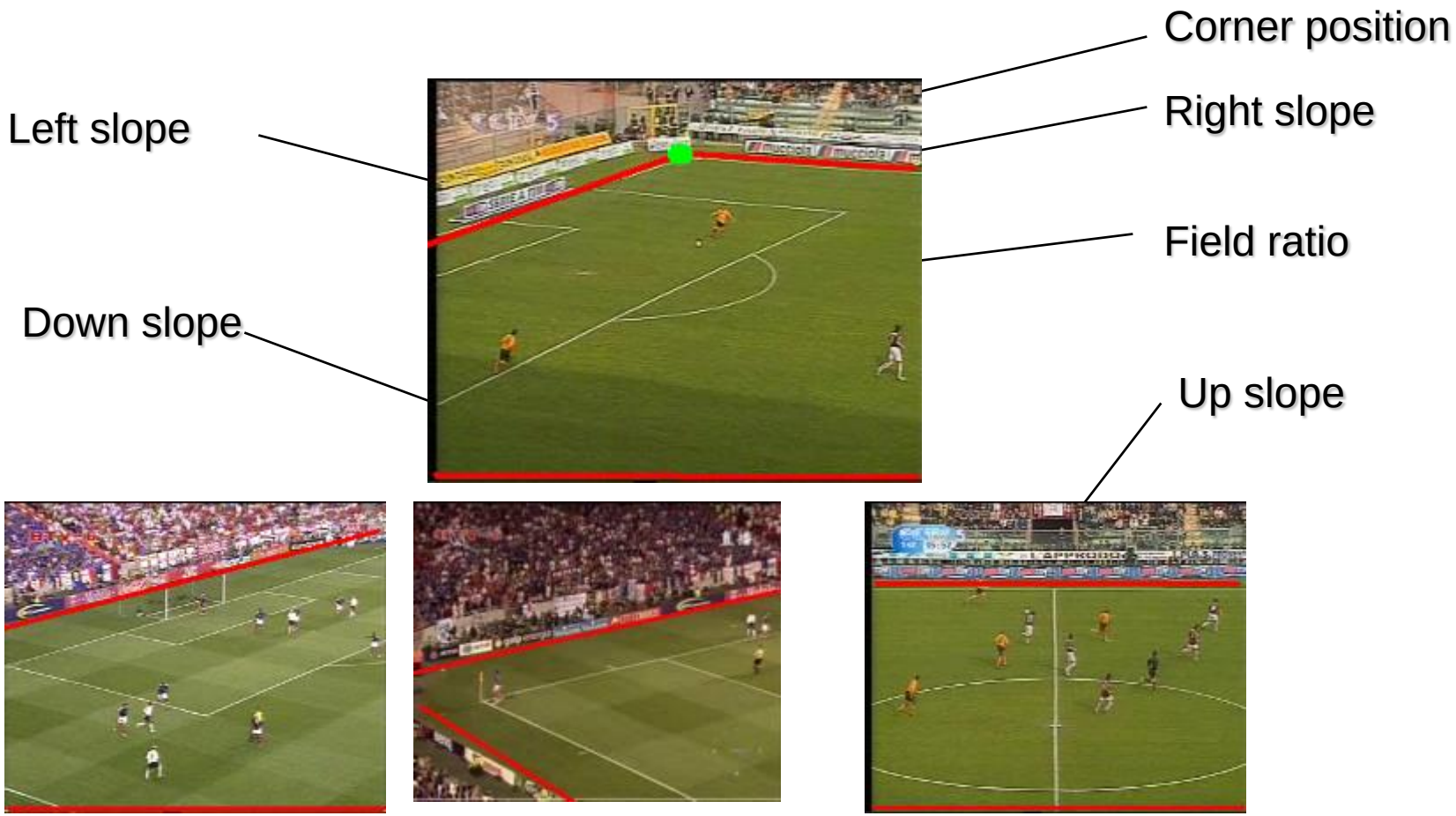
视频结构化：视角分类（足球）

- View classification framework

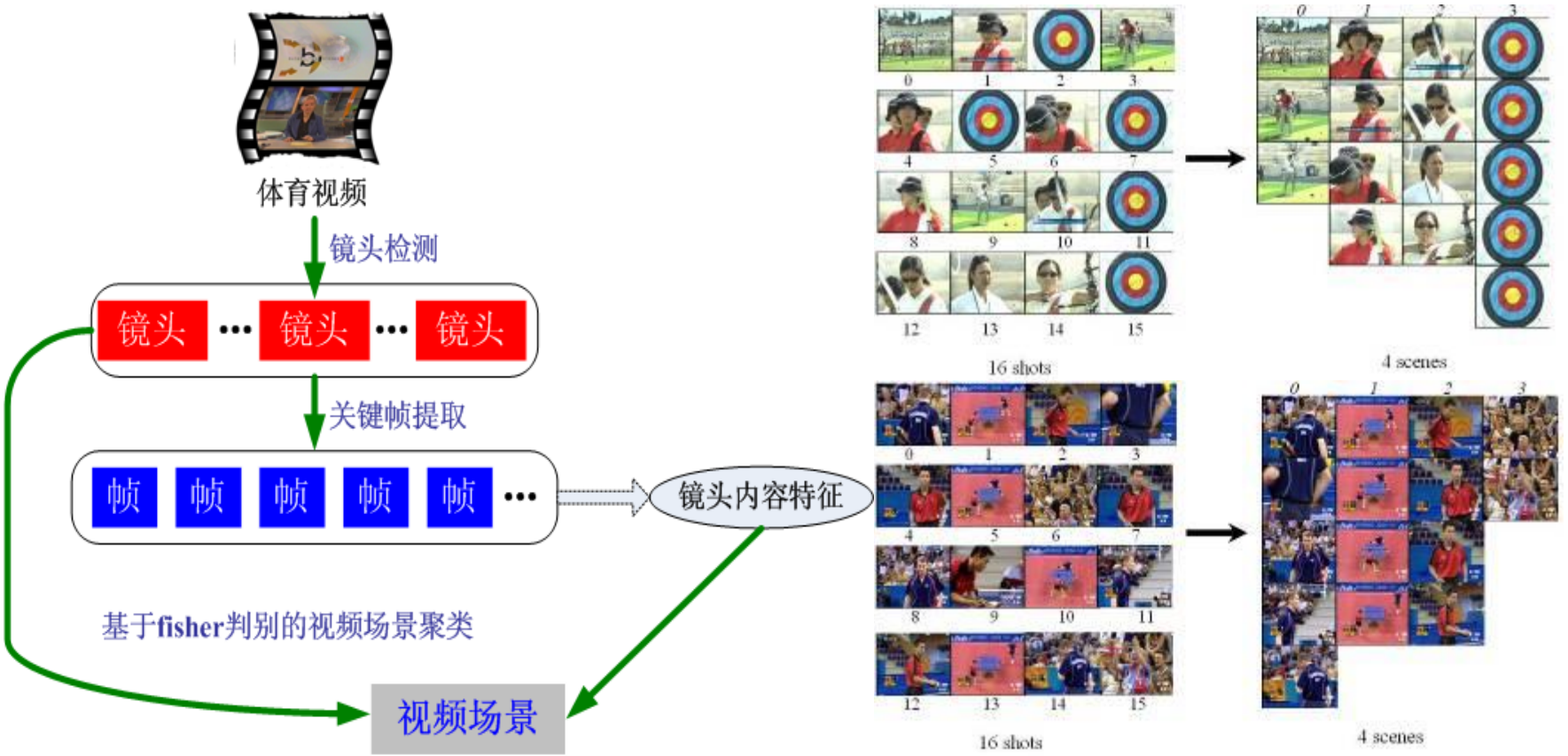


视频结构化：视角分类（足球）

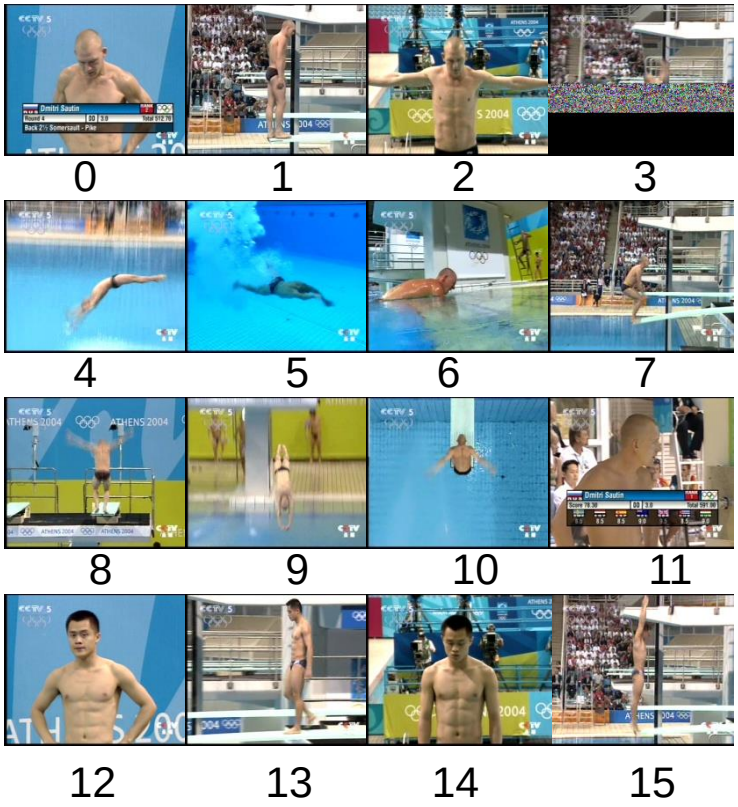
- Features for global view classification



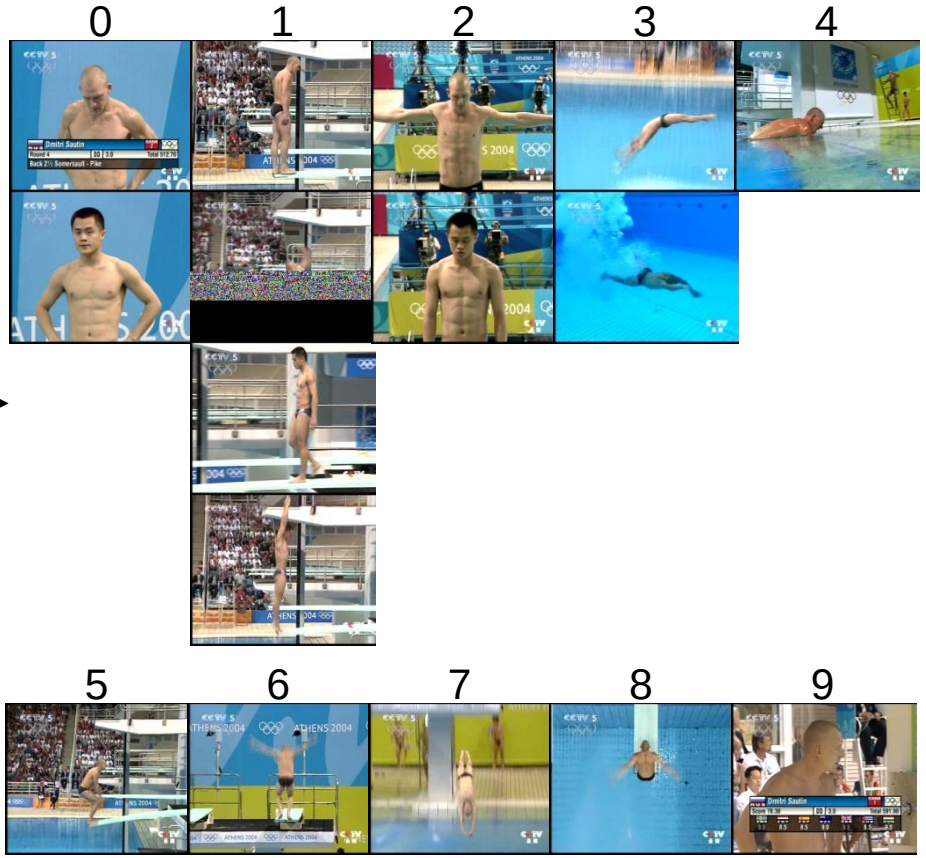
视频结构化：场景聚类


体育视频
镜头检测
镜头
镜头
镜头
关键帧提取
帧
帧
帧
帧
帧
镜头内容特征
基于fisher判别的视频场景聚类
视频场景
16 shots
4 scenes
16 shots
4 scenes

视频结构化：场景聚类

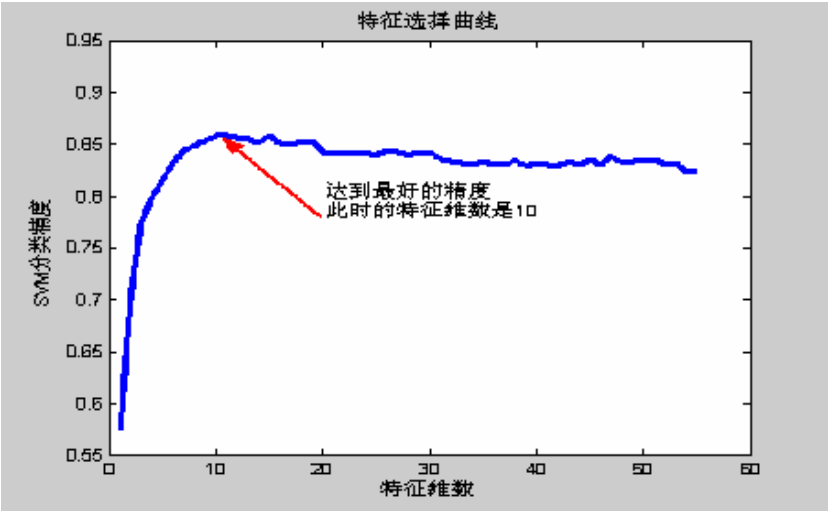


16 shots

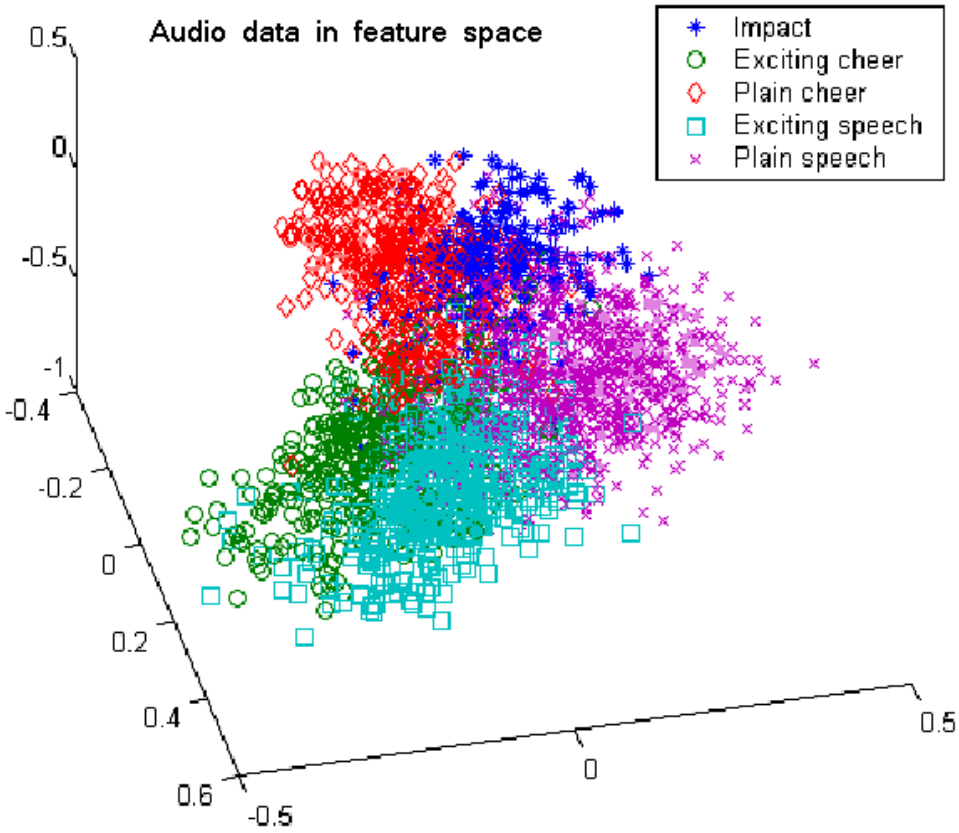


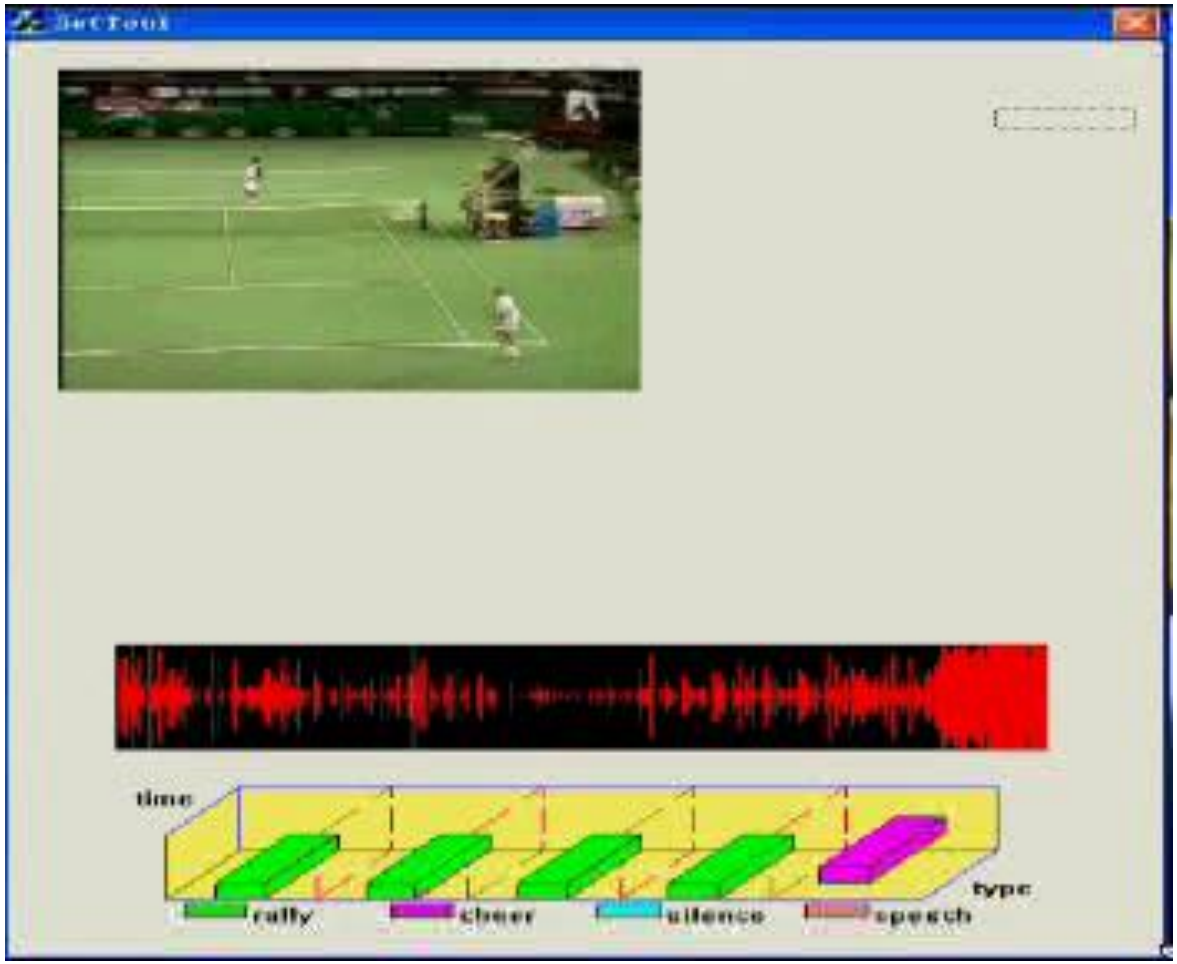
10 scenes

- **Four Audio Types**
  - **impact, cheer, speech, silence**
- **Features**
  - **MFCC, STE, Zero Crossing Rate, Brightness & bandwidth, Spectrum flux (55 dimensions total)**
  - **Optimal features selection based on forward search method**



- Classification
  - Samples preparation
    - Hundreds of training samples
    - Cross validation
  - GMM vs. SVM




time
type
rally
cheer
silence
speech

挥拍类型

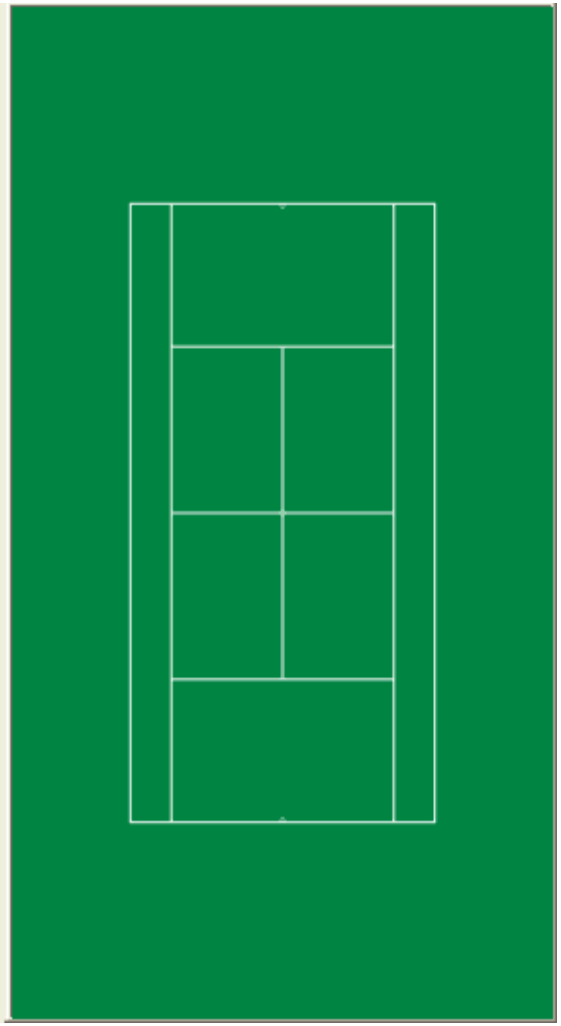
左挥拍: 0

右挥拍: 0

击球类型

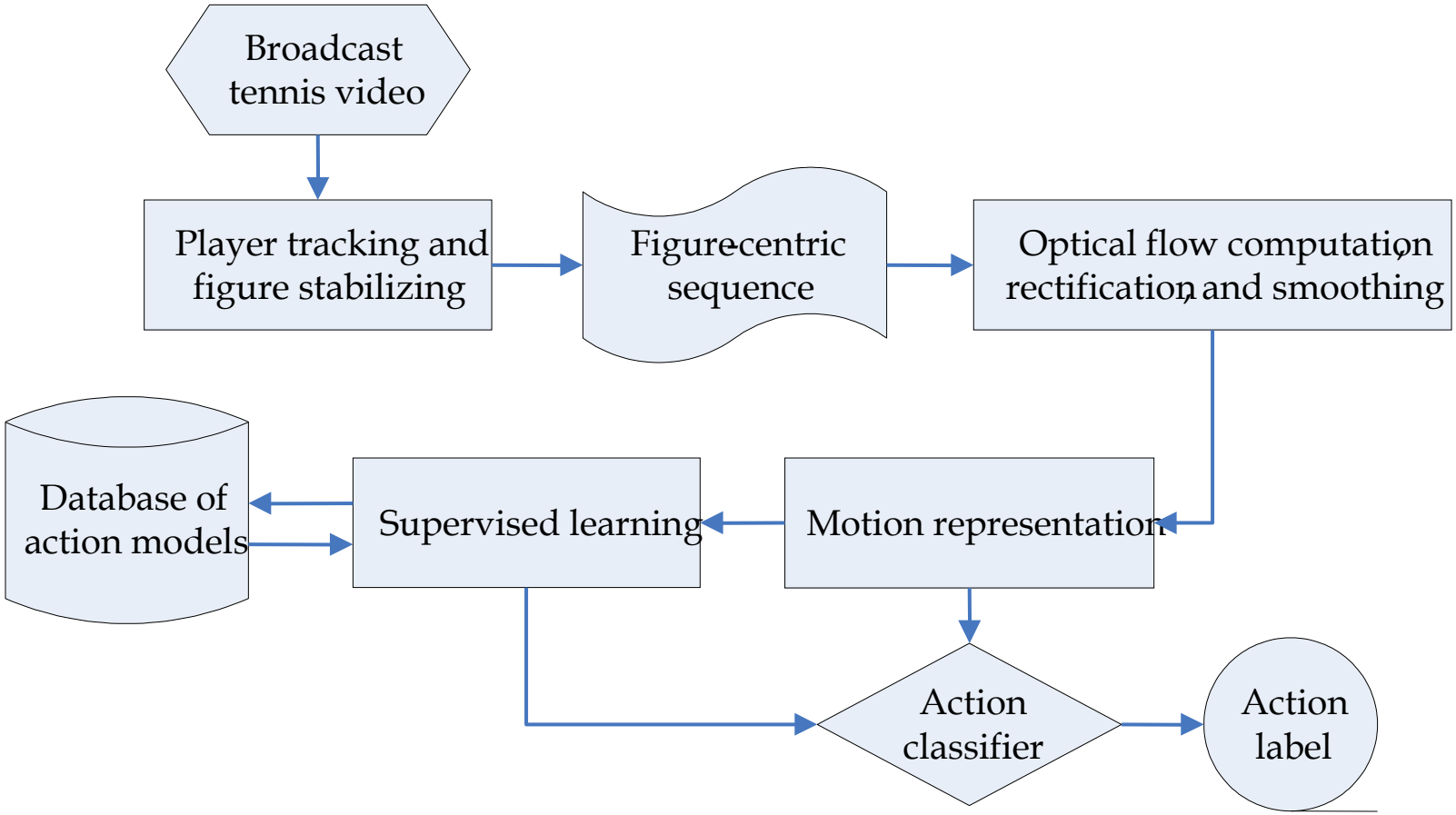
正手击: 0

反手击: 0

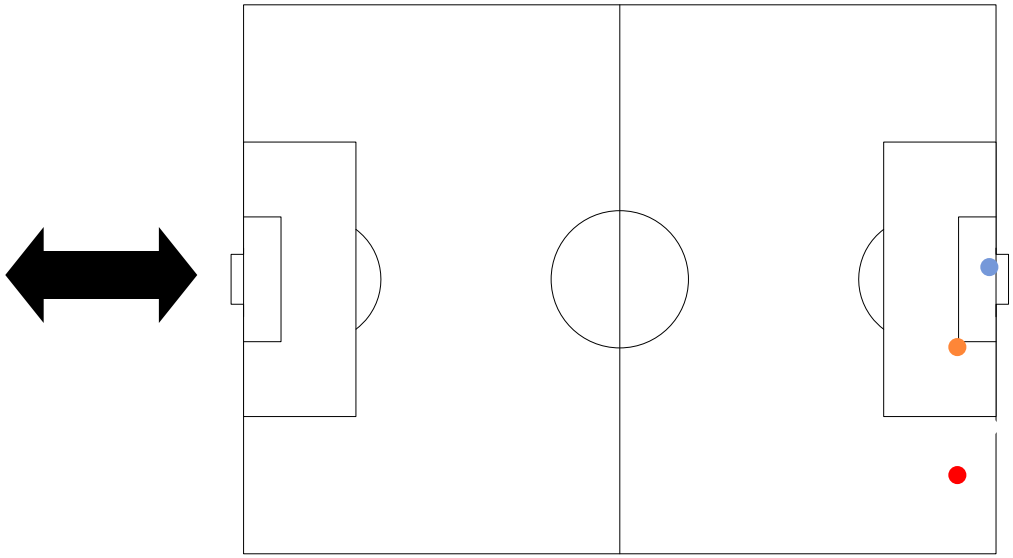


深层次理解：运动员活动识别

- Motion-based approach

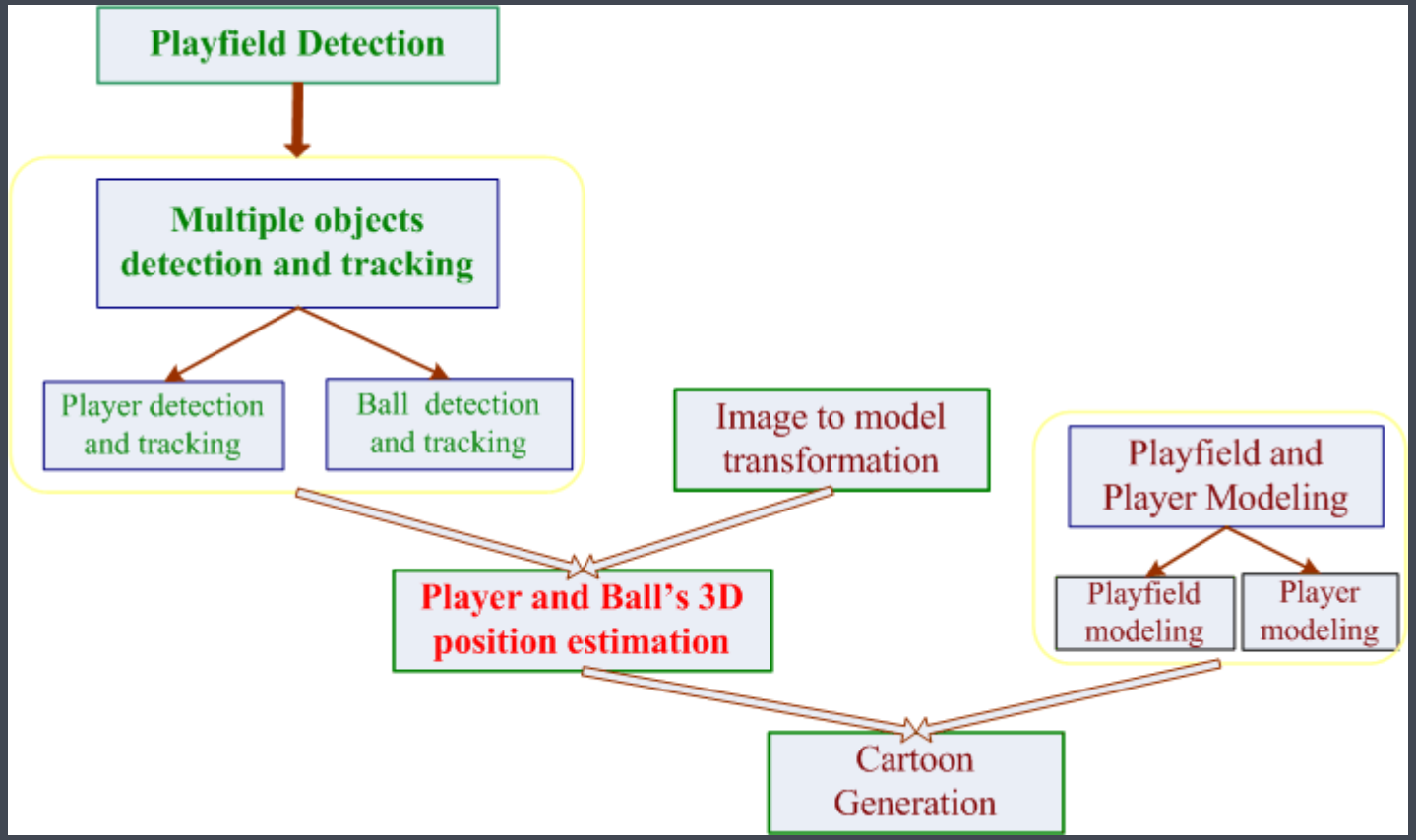


- Need at least four correspondence points, no three of them are collinear

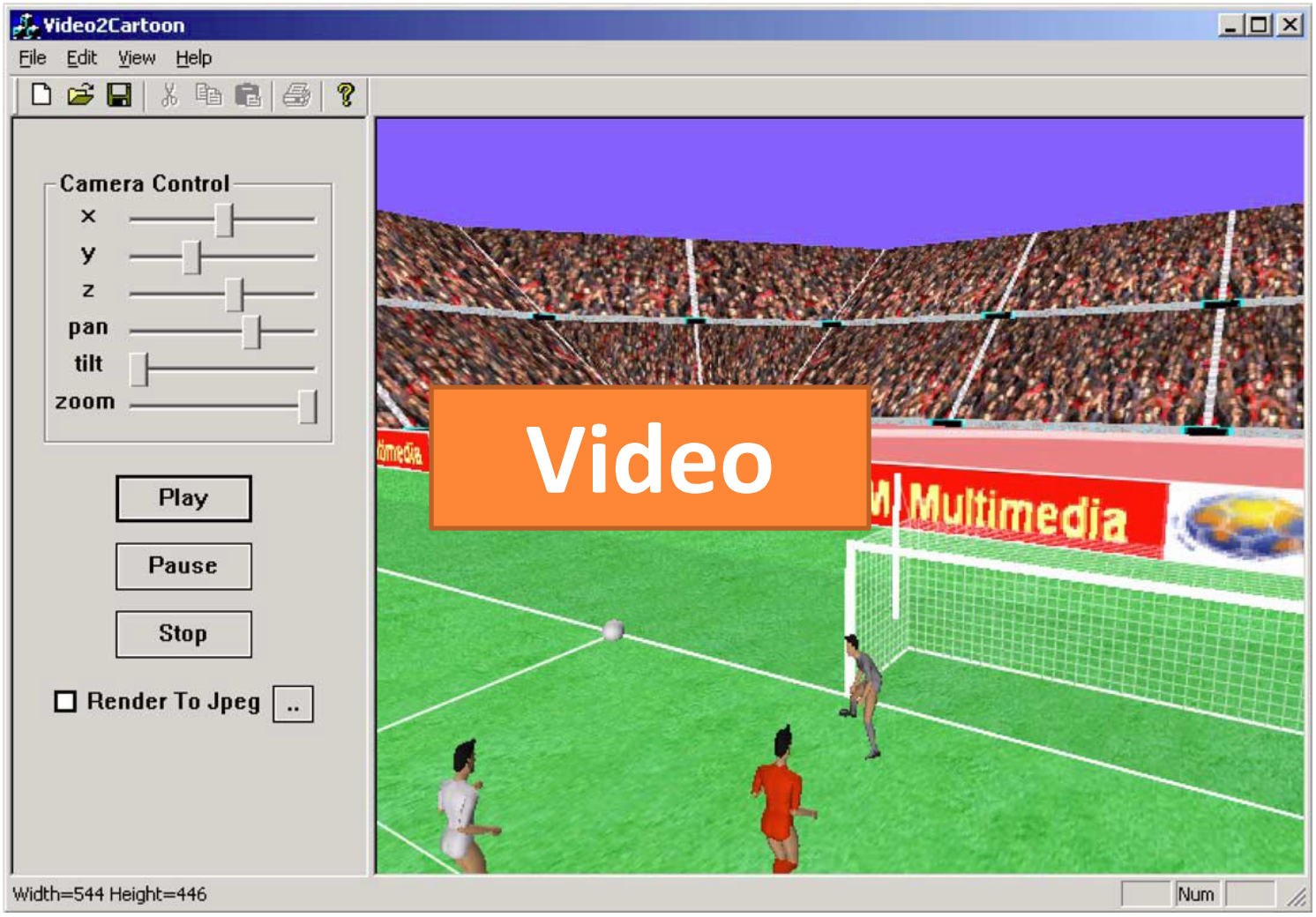


Yang Liu, Dawei Liang, Qingming Huang and Wen Gao, Extracting 3D information from broadcast soccer video, Journal of Image and Vision Computing, vol.24, no.10, pp1146-1162,Oct.2006

深层次理解：精彩片段三维重建



精彩片段三维重建


Video2Cartoon
File Edit View Help
Camera Control
x
y
z
pan
tilt
zoom
Play
Pause
Stop
Render To Jpeg
..
Video
Multimedia
Width=544 Height=446
Num

音频分析与检索

- Audio retrieval
  - Find required sound segment from audio database or broadcasting
  - Find interesting music from song/music database or web
- Methods of audio retrieval
  - Physical features of audio signal
    - 底层特征，可以直接从音频文件中计算得到，这些特征包括：响度、谱功率、亮度、带宽、音高、倒谱
  - Semantic features of audio
    - 语义特征是中高层特征，是从底层特征总结出来的，同底层特征相比，他们更加准确的反映了音频内容的特征。
      - Male/female, young/old
      - Rhythm and melody（节奏）
      - Timbre（音质、音色）

- 基于音频特征来对音频进行分类

1. 特征提取
   - **Reduce sound to small set of parameters**

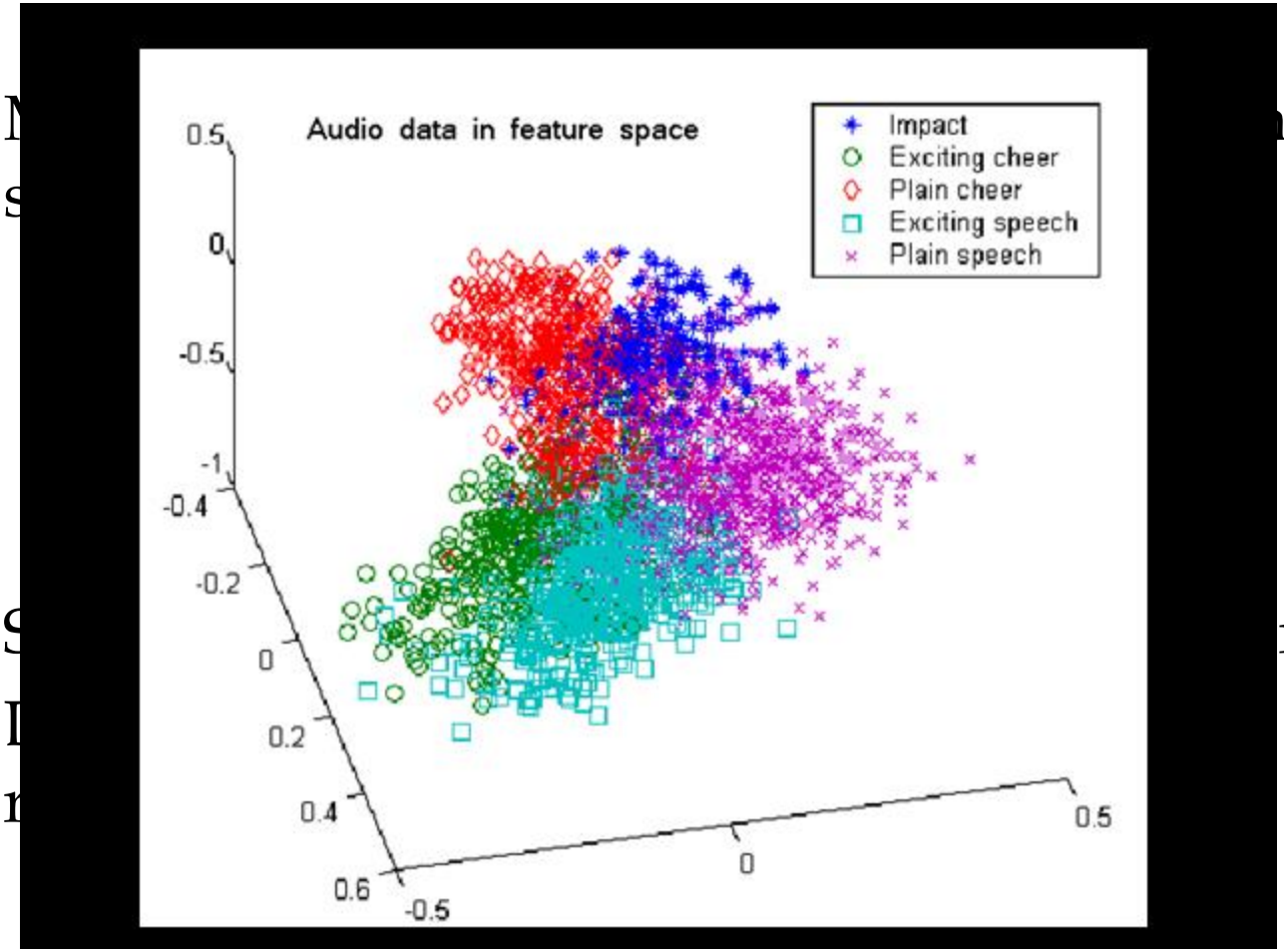
2. 分类方法
   - **To accomplish classification**

Acoustical Features

1. M… …n
s…

2. S… …r.

3. I…
r…



Acoustical Features: Loudness

- 通过短时傅里叶变换（STFT）计算得到的信号能量平方根来近似
- **Human ear: 120 db range**
- **Software: 100 db range from 16 bit recordings**

Acoustical Features: Pitch

- 音高是人语音波形的基本周期，是在语音信号分析中的一个重要参数。
- **Estimated by taking series of short-time Fourier spectra**
- **Human ear: 20Hz – 20kHz**

- Due to the difference in physiology, the pitch ranges for males and females are different:
  - The pitch range for males is 35 ~ 72 semitones, or 62 ~ 523 Hz.
  - The pitch range of females is 45 ~ 83 semitones, or 110 ~ 1000 Hz.

Acoustical Features: Brightness

- Measure of higher frequency content of signal
- Computed as centroid of the short-time Fourier magnitude spectra
- Stored as log frequency
- Varies over same range as pitch
- Can’t be less than pitch estimate at any given instant

Acoustical Features: Bandwidth

- 谱分量和质心差值的功率加权平均值。

1、声音：由振动而产生，通过空气进行传播。它由许多不同频率的谐波所组成，谐波的频率范围称为声音的带宽(bandwidth)，带宽是声音的一项重要参数。
2、多媒体技术处理的声音信号主要是人耳可听到的20～20kHz的音频信号(audio)
言语（speech）/语音：人说话的声音，其频率范围约为300～3400Hz
全频带声音：音乐声、风雨声、汽车声等其他声音，其带宽可达到20～20kHz

Feature Vector

- These aspects of sound vary over time. **Trajectory in time computed**
- **For each trajectory, computes & stores:**
  - **Average**
  - **Variance**
  - **Autocorr**
    - 对轨
  - **Duration**

Table 1. Male laughter. Duration: 2.12571.

Property	Mean	Variance	Autocorrelation
Loudness	-54.4112	221.451	0.938929
Pitch	4.21221	0.151228	0.524042
Brightness	5.78007	0.0817046	0.690073
Bandwidth	0.272099	0.0169697	0.519198

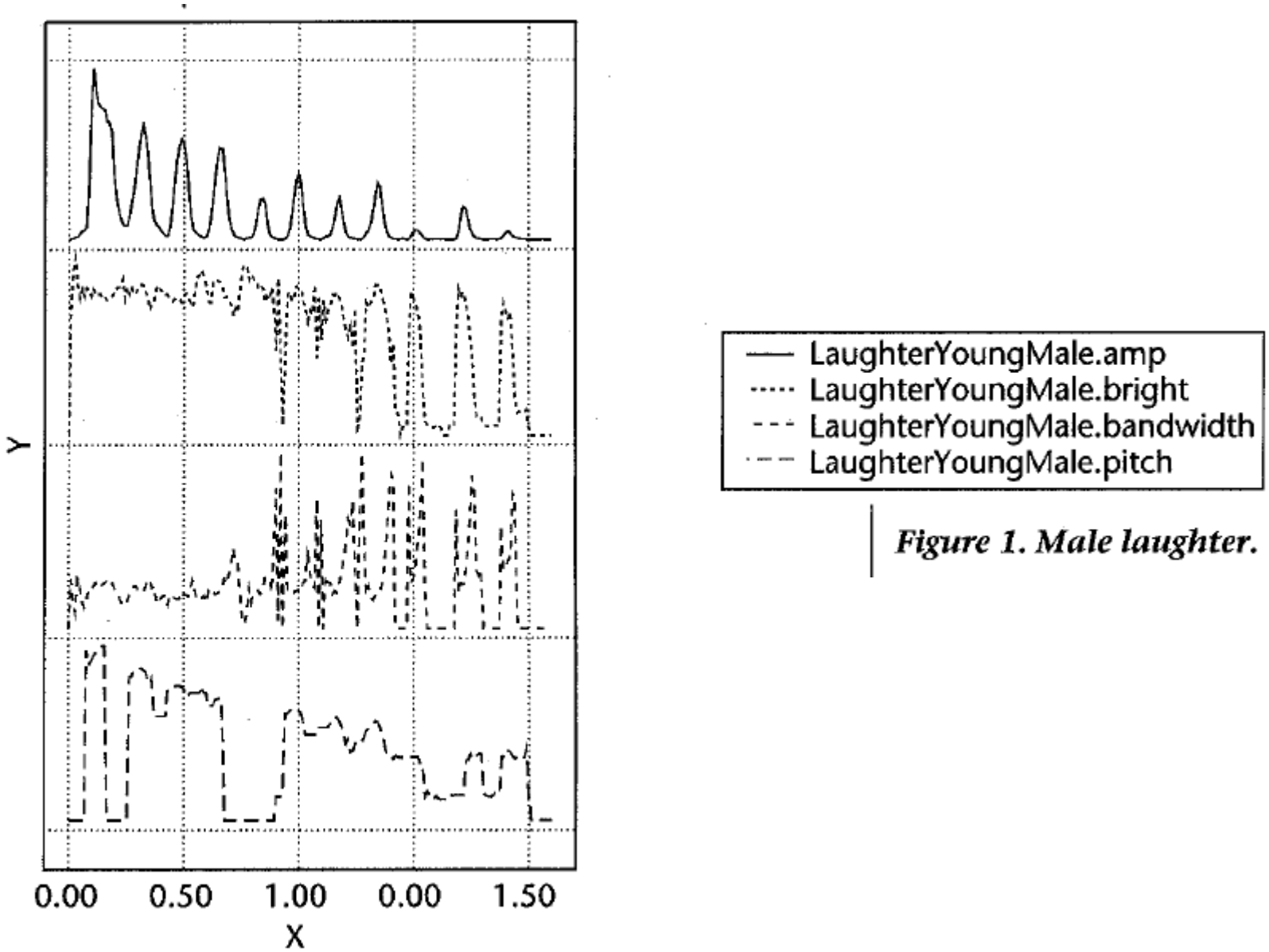


Figure 1. Male laughter.

Training The System

- For each sound entered into the db, the *N*-vector, *a*, is computed
- Mean vector and covariance matrix *R* for the *a* vectors in each class are calculated:

$$\mu = (1/M) \sum_j .a[j]$$

$$R = (1/M) \sum_j .(a[j]-\mu)(a[j]-\mu)T$$

- Mean + Covariance = System's model of perceptual property being trained by user

Classifying Sounds

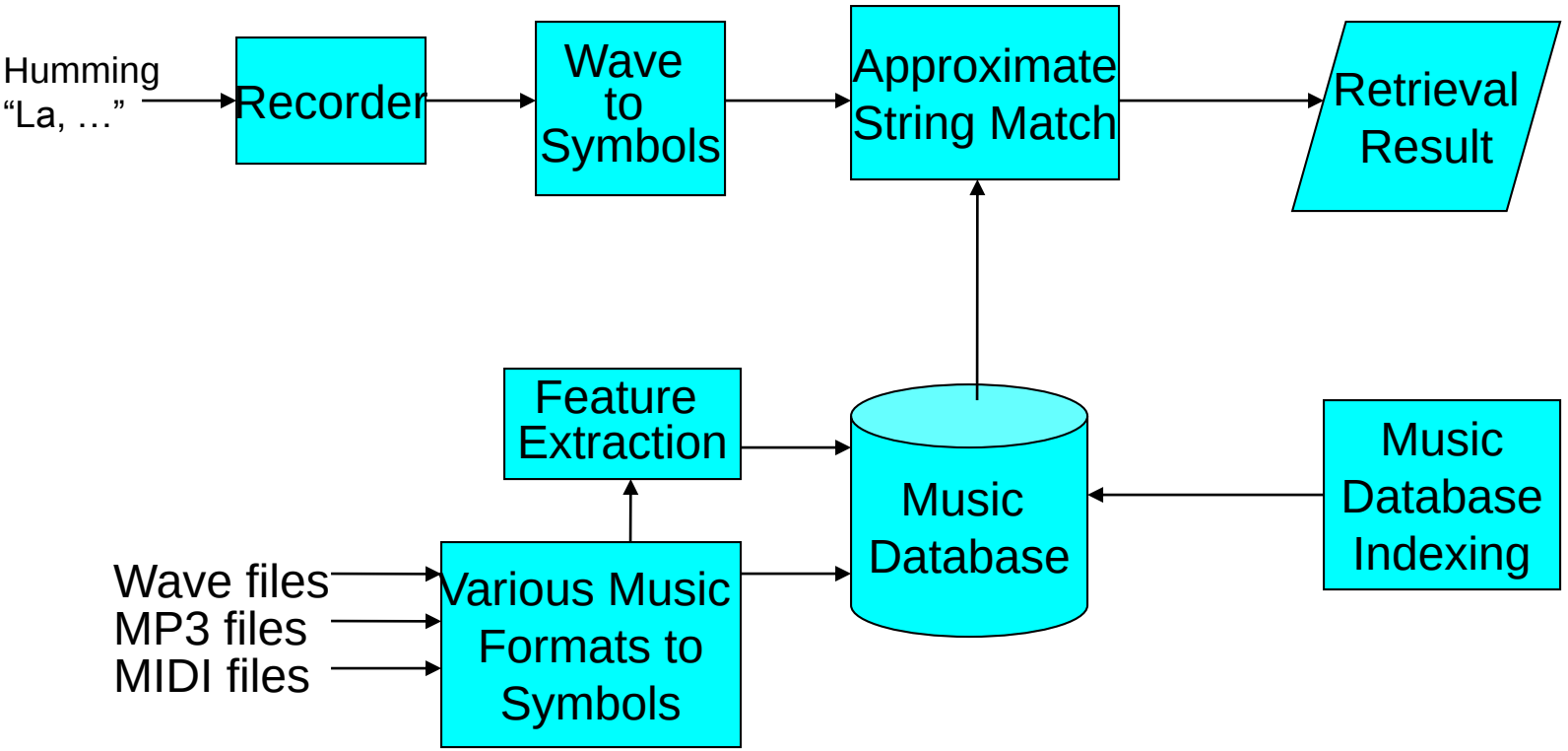
- When a new sound needs to be classified, a distance measure is calculated from new sound's *a* vector and previous model
- Using weighted L_2 or Euclidean distance:

$$D = ((a-\mu)^T R^{-}1(a-\mu))^{1/2}$$

- Likelihood value *L* based on normal distribution and given by:

$$L = exp(-D^2/2)$$

Music Retrieval by Singing/humming



多媒体分析与检索技术

- 多媒体检索概论
- 基于内容的图像分析与检索（CBIR）
- 视频分析与检索
- 音频分析与检索

  - Framework
  - CBIR
  - Semantic Gap
  - CBIR Features
  - Video Features
  - Video Structure…