机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

信息与通信工程学院 网络搜索教研中心

北京邮电大学

引子：从身边的学习说起…

- 思考一下我们日常生活中的学习过程
  - 什么是学习
    - 学习: 获取知识、形成技能，获得适应环境、改变环境的能力的过程
  - 学习是怎样一个过程
    - 通过了解、认识、记忆、理解“问题与答案”来进行，进而达到能够自由运用与推广
  - 怎样评价学习的效果
    - 测试
      - 测试题目是原封不动所学过的：考察记忆能力
      - 测试题目是学过的但改变了形式的: 考察运用与推广能力
  - 怎样有效地学习

机器学习的定义

A computer program is said to learn from experience E

with respect to some tasks T and

performance measure P,

if its performance at tasks in T,

as measured by P,

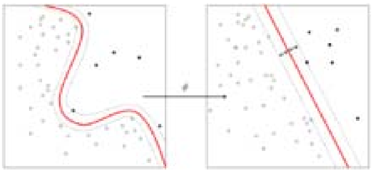
improves with experience E.

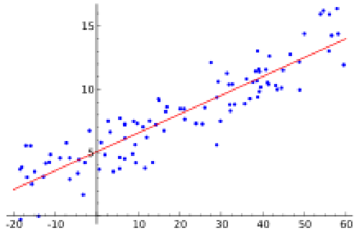
E: 数据　　P: 性能评价指标　　T: 特定任务

概念辨析

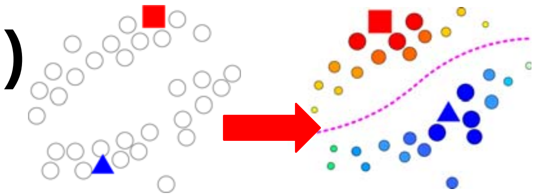
- 数据挖掘
  - 源于数据仓库技术
  - 数据库中的知识发现(Knowledge-Discovery in Databases)
- 模式识别
  - 源于自动控制工程，跟人的识别能力相对应的各种应用问题及其相关联的一系列技术
- 机器学习
  - 人工智能的偏于理论的分支，从具体应用问题抽象出来的具有一般性的模型、算法与理论
- 数据科学
  - 与数据的感知、获取、处理和分析相关的模型、算法和理论
    - 对应于机器学习中的无监督部分，又区别于经典信号处理与分析
    - 区别与联系: 信号处理、应用数学、量子物理

学习问题的基本形式





- 按照是否有监督信息使用
  - 有监督学习(Supervised learning)
    - 分类: 输出为离散变量
    - 回归: 输出为连续变量
  - 无监督学习(Unsupervised learning)
    - 聚类: 寻找数据的某种划分
    - 降维: 寻找数据的“坐标系”变换
  - 半监督学习(Semi-supervised learning)



- 表示符号说明
  - 训练数据(training data)

$$(X,Y)=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_n,y_n)\}$$

  - 数据本身

$$X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$$

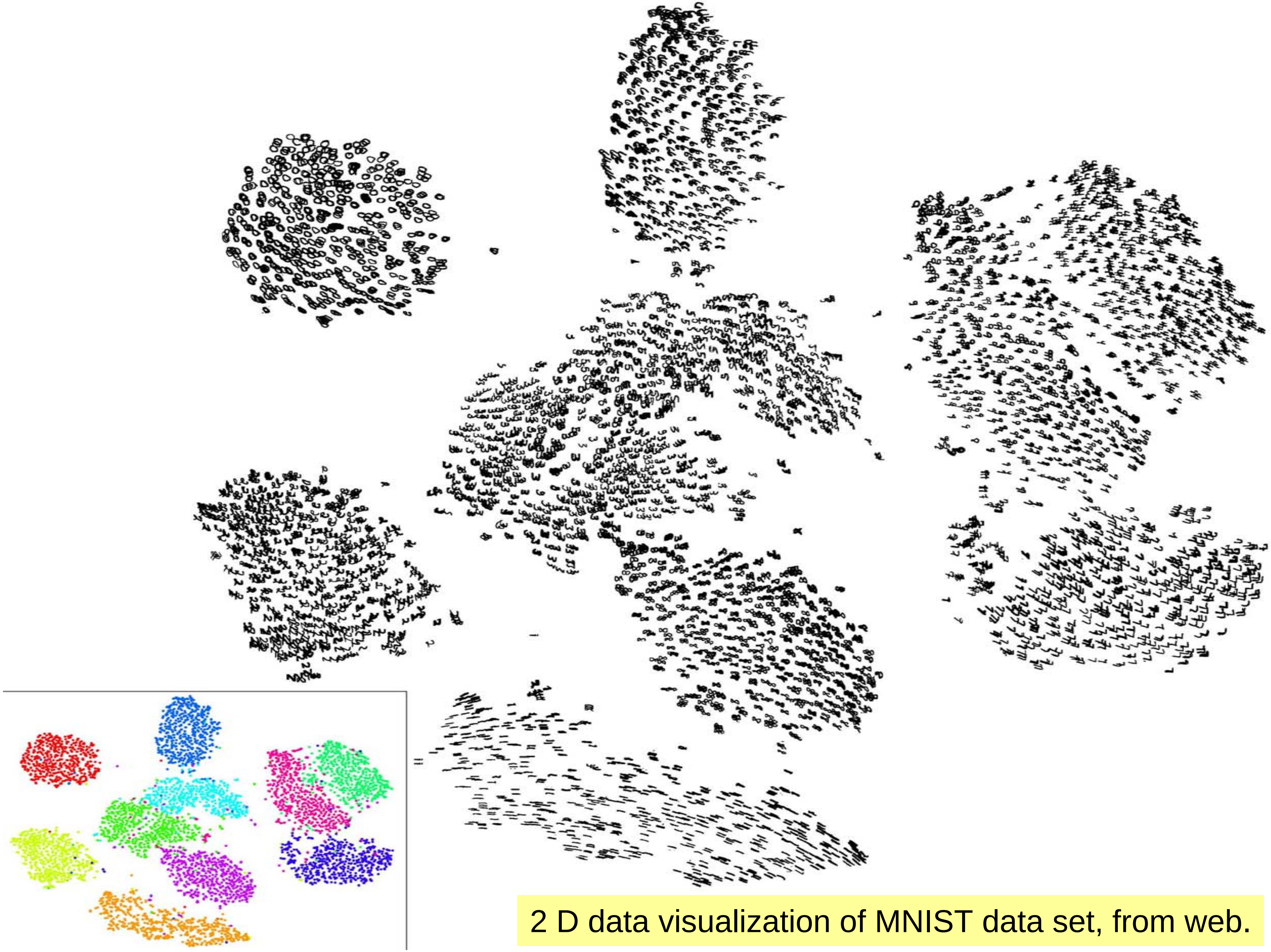
应用问题举例：I

- 手写数字识别
- 手写汉字识别
- 人脸识别
- 材料、树叶、花卉、食物识别
- 自然场景分类
- 图像物体归类
- 动态场景识别
- 文本分类
- 文本倾向性判断

手写数字图像识别

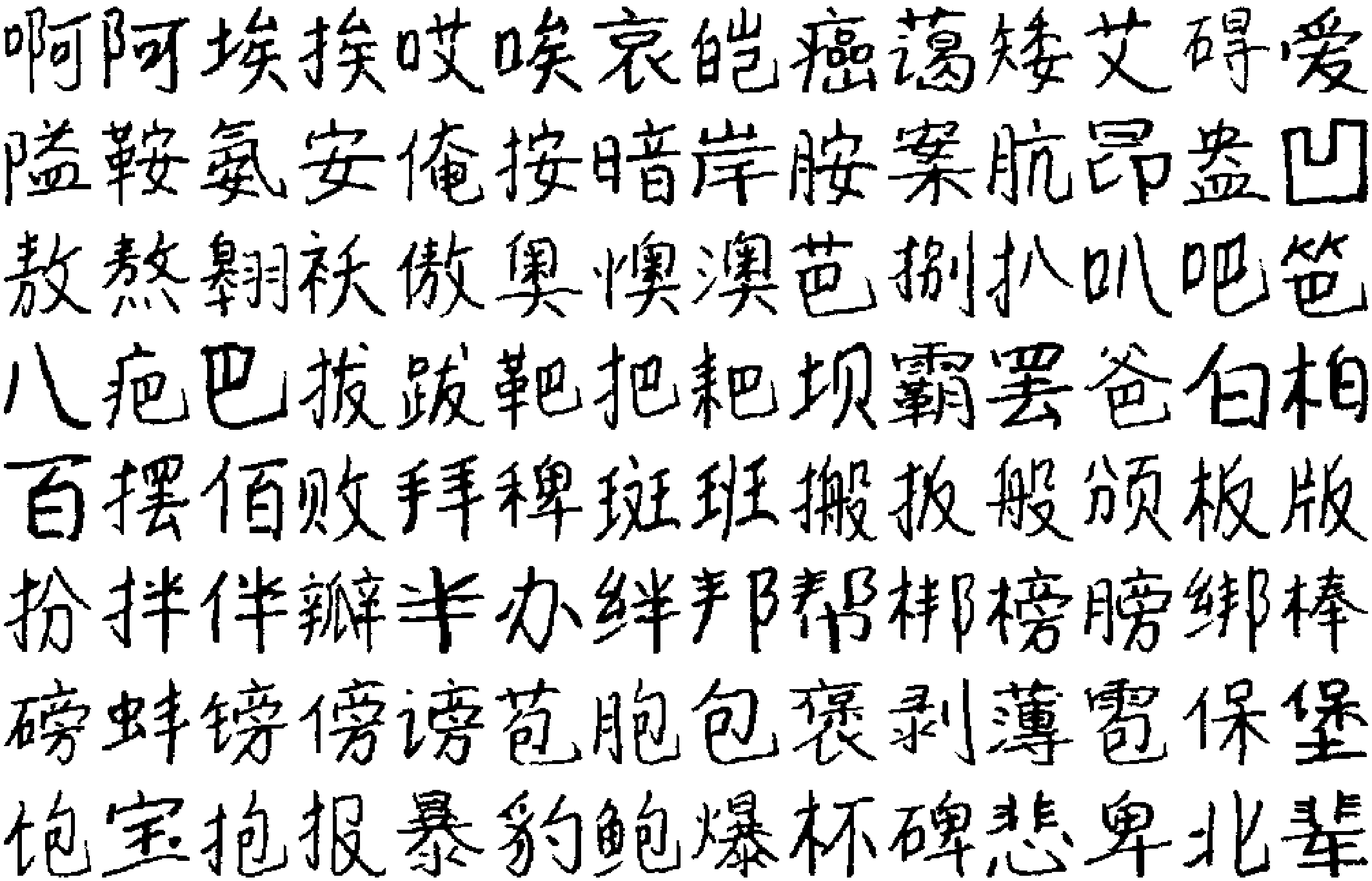
-

from web



2 D data visualization of MNIST data set, from web.

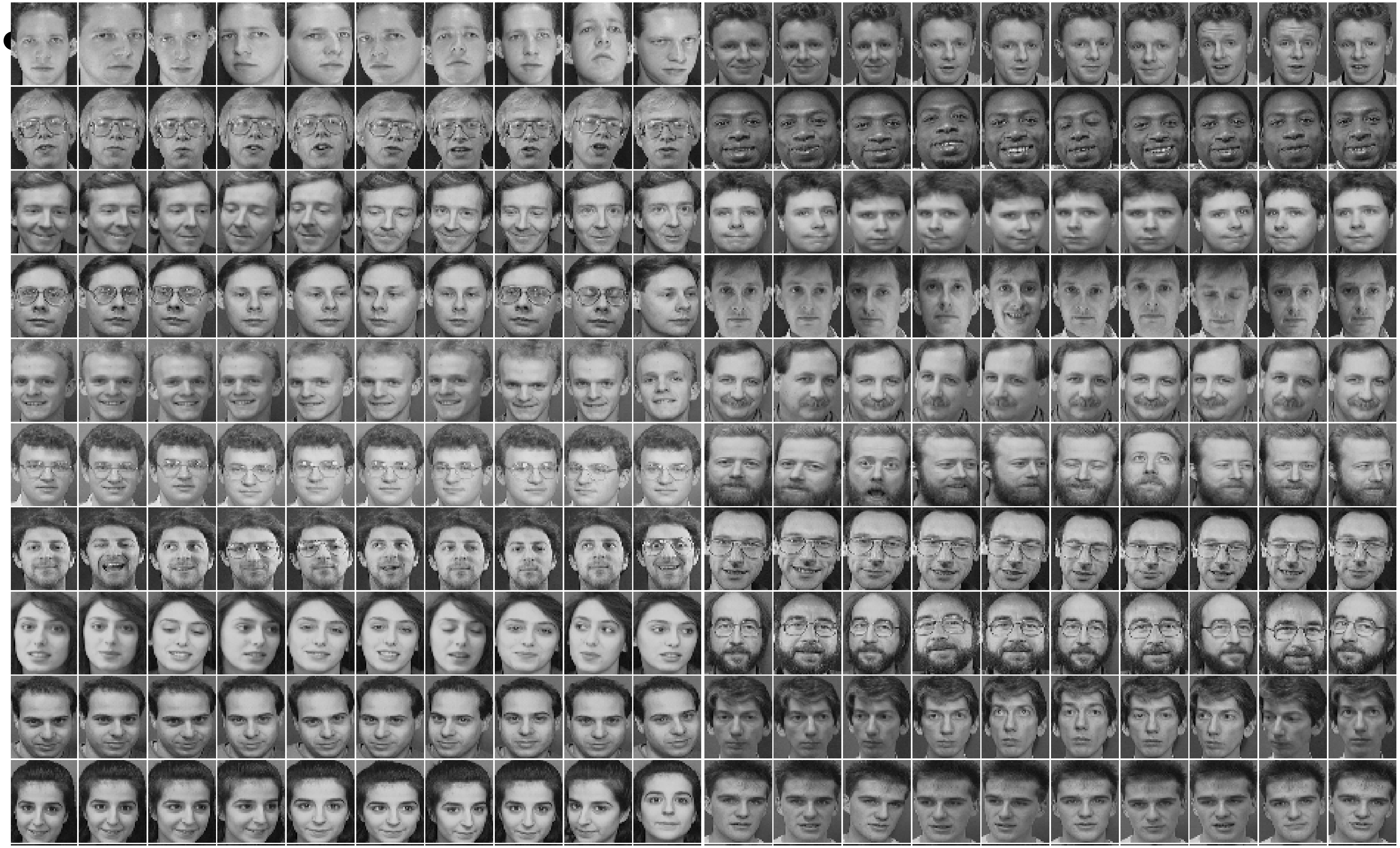
手写汉字图像识别

- 

啊阿埃挨哎唉哀皑癌蔼矮艾碍爱
隘鞍氨安俺按暗岸胺案肮昂盎凹
敖熬翱袄傲奥懊澳芭捌扒叭吧笆
八疤巴拔跋靶把耙坝霸罢爸白柏
百摆佰败拜稗斑班搬扳般颁板版
扮拌伴瓣半办绊邦帮梆榜膀绑棒
磅蚌镑傍谤苞胞包褒剥薄雹保堡
饱宝抱报暴豹鲍爆杯碑悲卑北辈

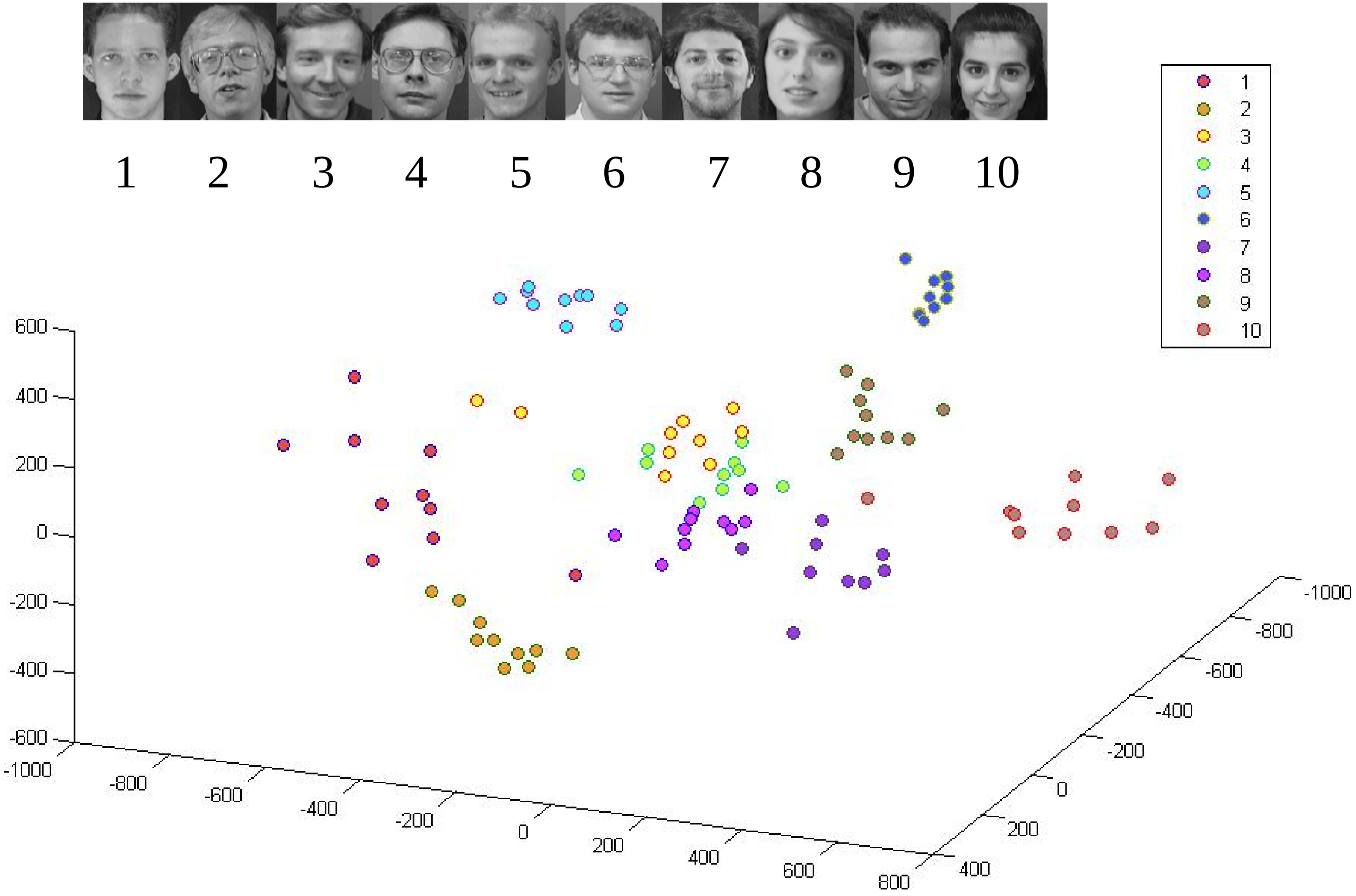
[1] H. Zhang et al., "HCL2000 -- A Large-scale Handwritten Chinese Character Database for Handwritten Character Recognition", ICDAR 2009, pp.286-290

人脸识别 (from ORL)



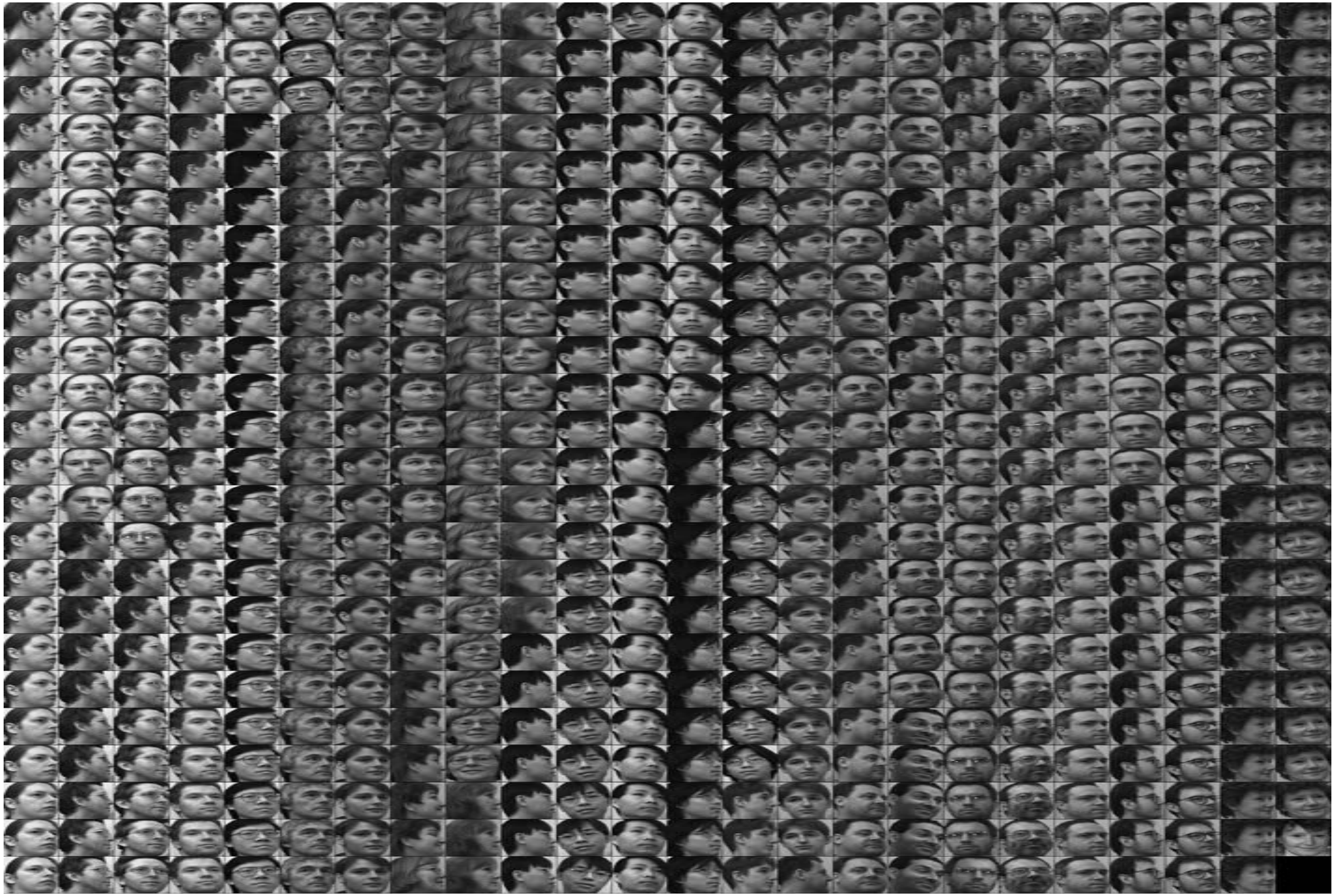
from ORL data set

ORL 数据集的可视化(by PCA)

- 

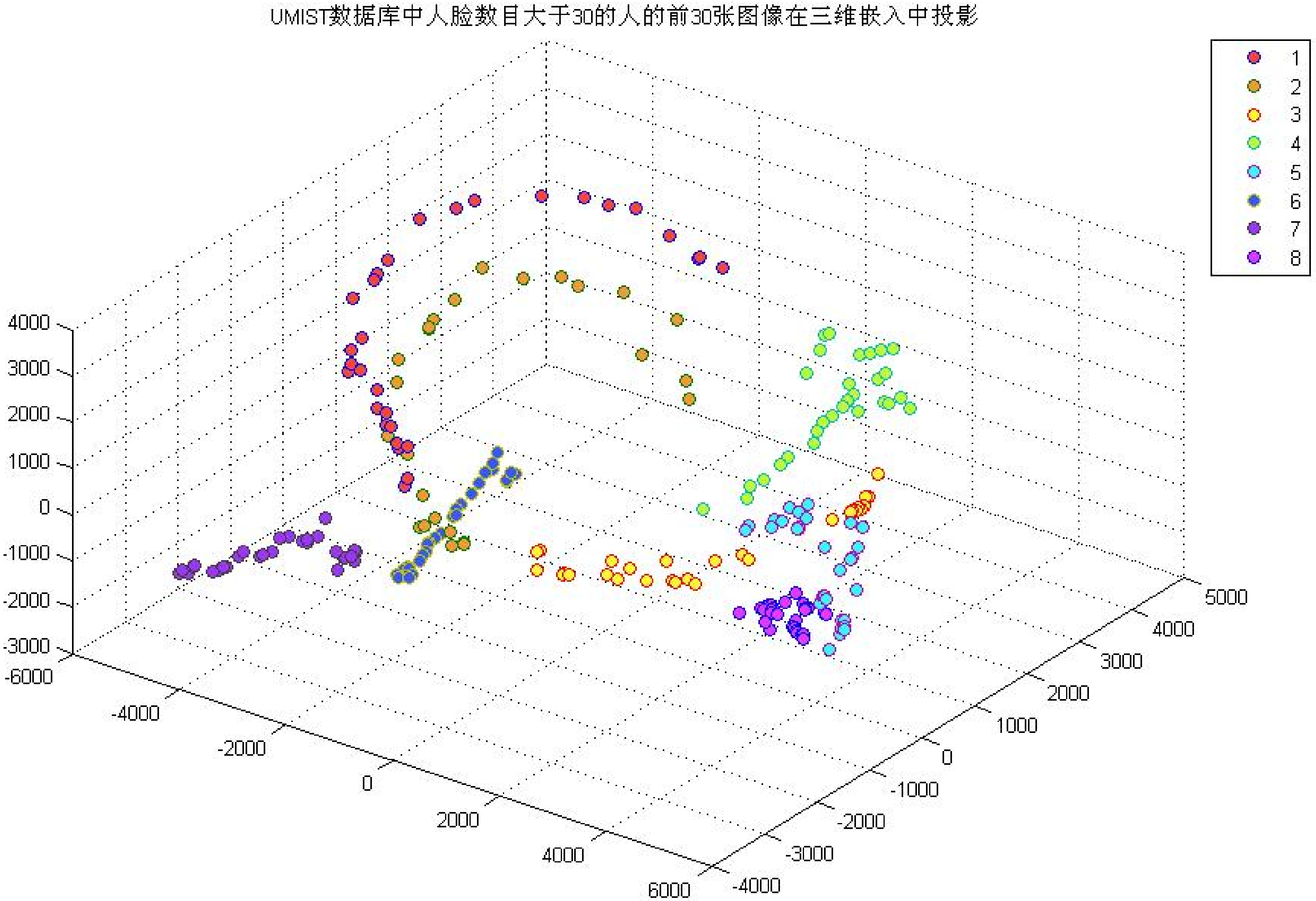
前10个人在前3维特征脸中的投影分布

人脸识别 (from UMIST)

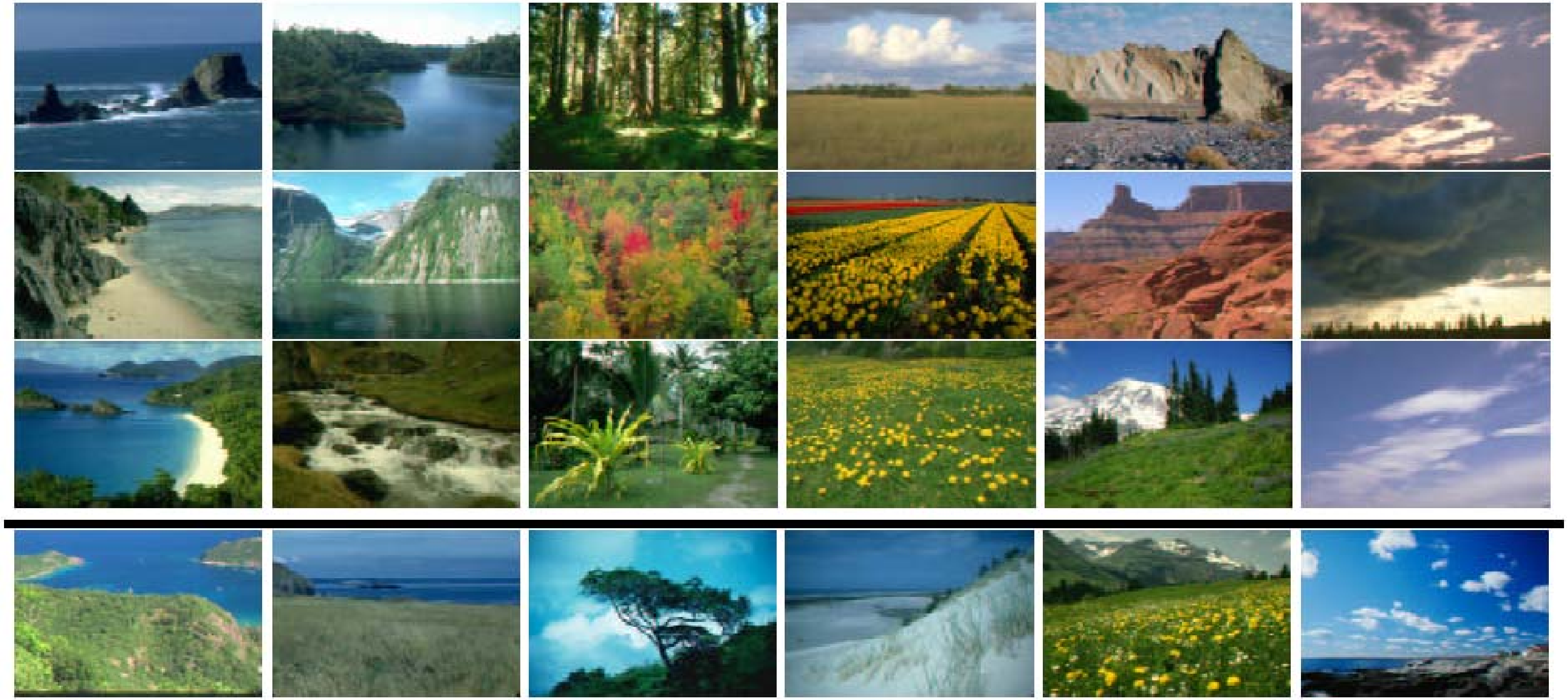
- 

UMIST 数据集的可视化(by ISOMAP)

-



自然场景图像分类



如何获得向量化表达?

From Dataset Scene-15

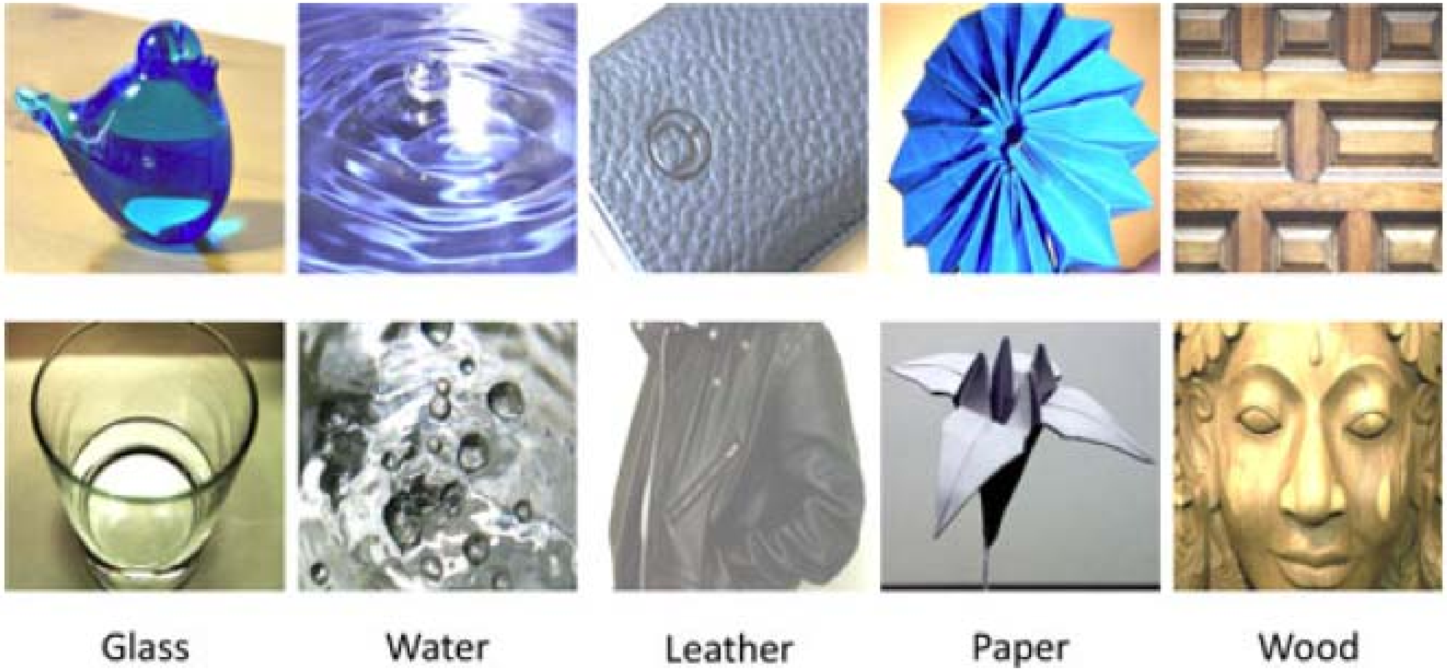
图像物体分类(VOC)

From Dataset Caltech 256

如何获得向量化表达?

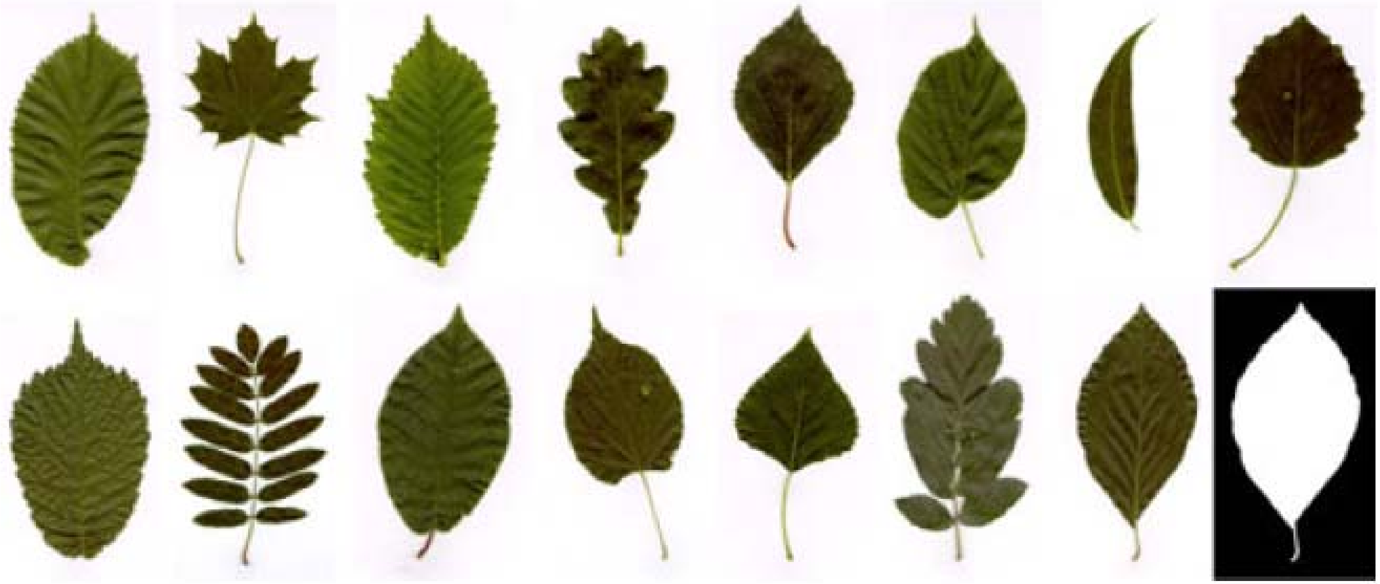
材料识别

-



[2] Xianbiao Qi, Rong Xiao, Chun-Guang Li, Yu Qiao, Jun Guo, and Xiaoou Tang, “Pairwise Rotation Invariant Co-occurrence Local Binary Pattern”, IEEE TPAMI, Vol. 36, No. 11, Nov. 2014, pp.2199-2213.

树叶识别

- 

[2] Xianbiao Qi, Rong Xiao, Chun-Guang Li, Yu Qiao, Jun Guo, and Xiaoou Tang, “Pairwise Rotation Invariant Co-occurrence Local Binary Pattern”, IEEE TPAMI, Vol. 36, No. 11, Nov. 2014, pp.2199-2213.

花卉识别

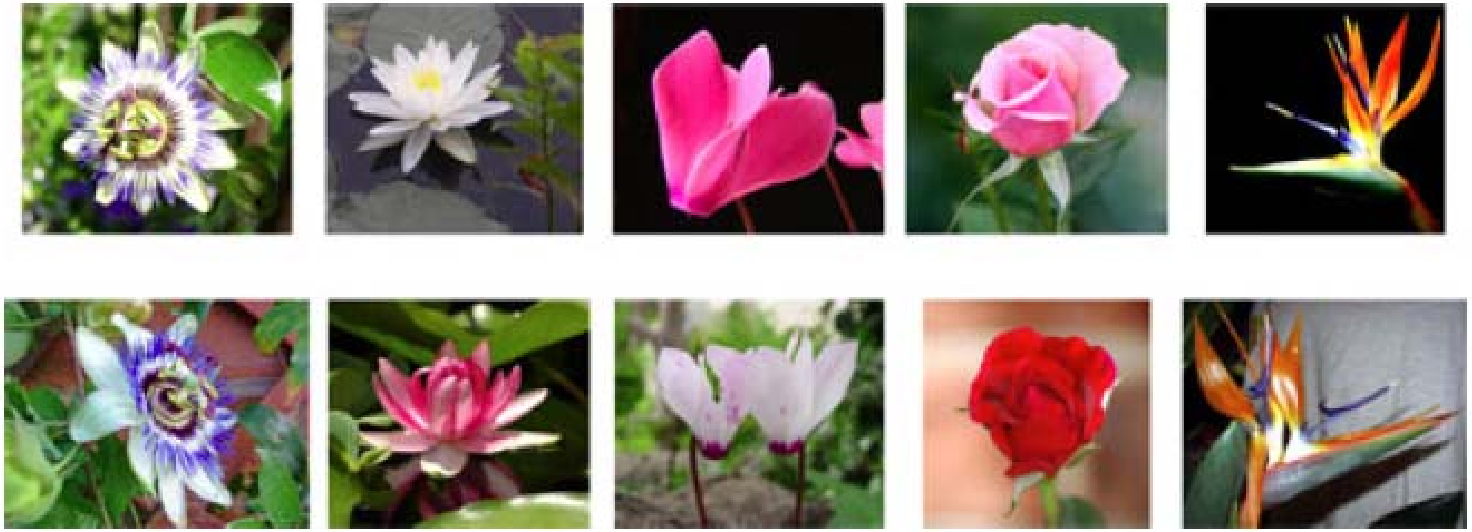
- 

Fig. 5. Some sample from Oxford Flower 102 data set. (The two images in each column from the same category).

[2] Xianbiao Qi, Rong Xiao, Chun-Guang Li, Yu Qiao, Jun Guo, and Xiaoou Tang, “Pairwise Rotation Invariant Co-occurrence Local Binary Pattern”, IEEE TPAMI, Vol. 36, No. 11, Nov. 2014, pp.2199-2213.

食物识别

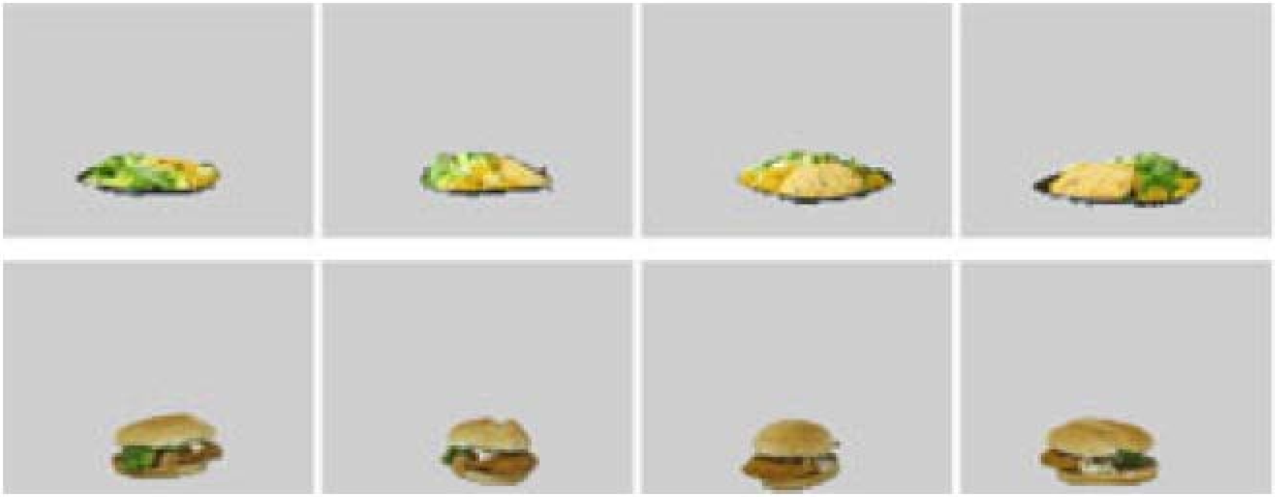


Fig. 7. Sample images from PRID food data set.

[2] Xianbiao Qi, Rong Xiao, Chun-Guang Li, Yu Qiao, Jun Guo, and Xiaoou Tang, “Pairwise Rotation Invariant Co-occurrence Local Binary Pattern”, IEEE TPAMI, Vol. 36, No. 11, Nov. 2014, pp.2199-2213.

动态纹理 (from YUPENN)

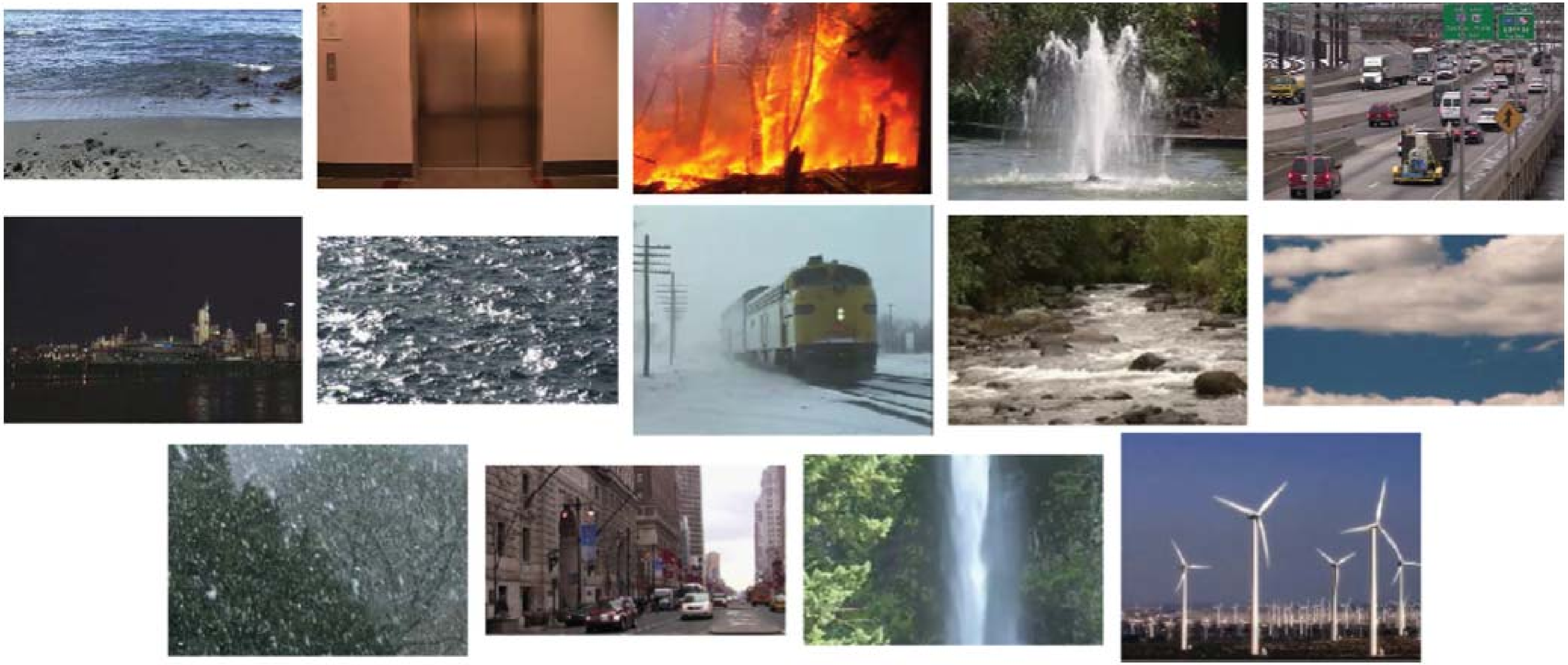


Fig. 6. Sample frames from dynamic scene data set YUPENN. Each image corresponds to a category of video sequence.

[3] Xianbiao Qi, Chun-Guang Li, Guoying Zhao, Xiaopeng Hong, Matti Pietikainen, “Dynamic texture and scene classification by transferring deep image features”, Neurocomputing, Vol.171, 2016, pp:1230-1241.

如何构造一个学习问题？

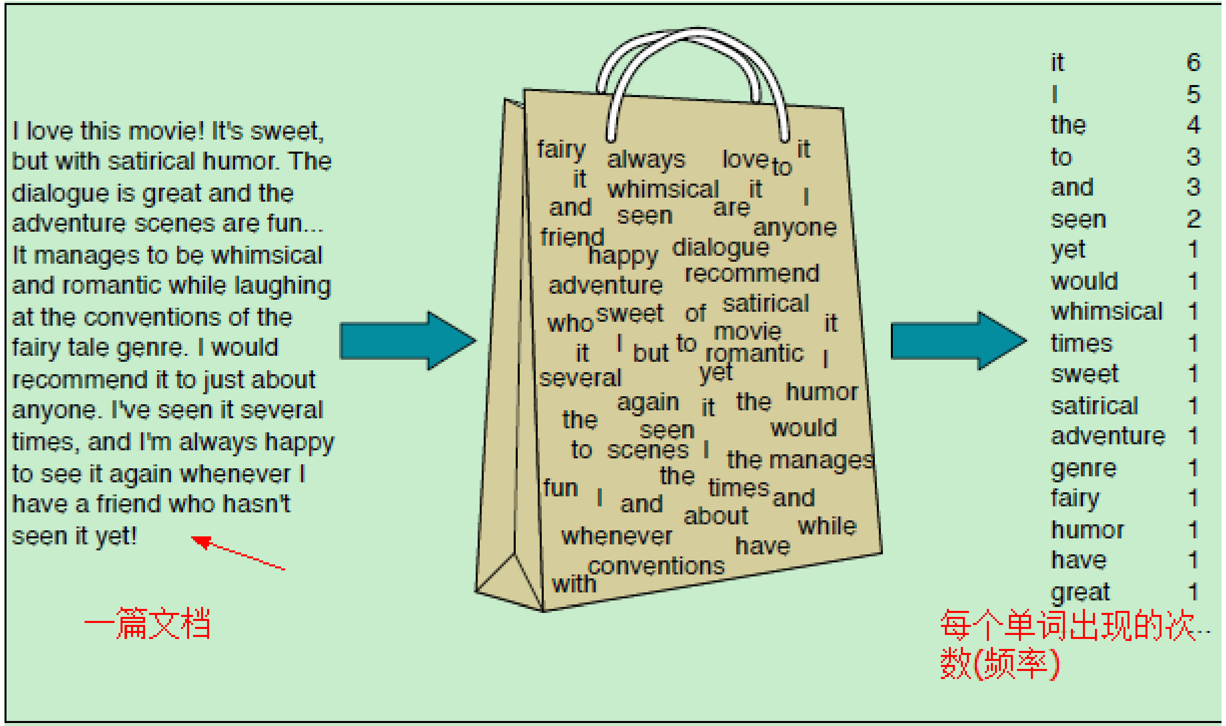
- 收集数据
  - 建立数据的表示，有时也叫做特征抽取或学习

- 明确特定任务，建立任务的抽象模型
  - 分类 **/** 回归 **/** 推荐 **/ Ranking /** 聚类 **/ …**

- 定义性能评价指标
  - 误差 **/** 错误率 **…**

应用问题举例 1: 文本分类

- 基于词袋模型建立直方图特征向量
  - Bag of Words



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

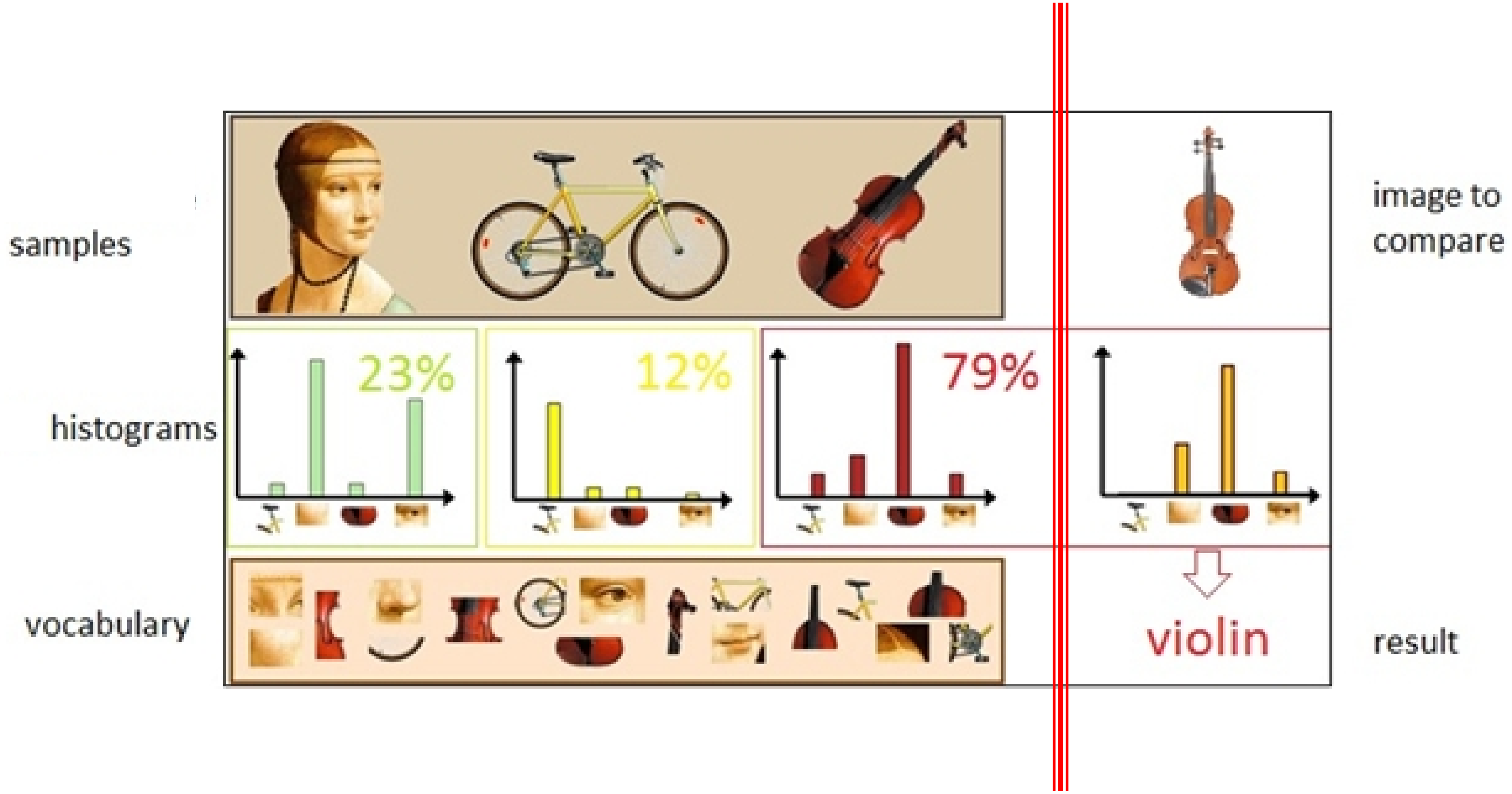
一篇文档

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

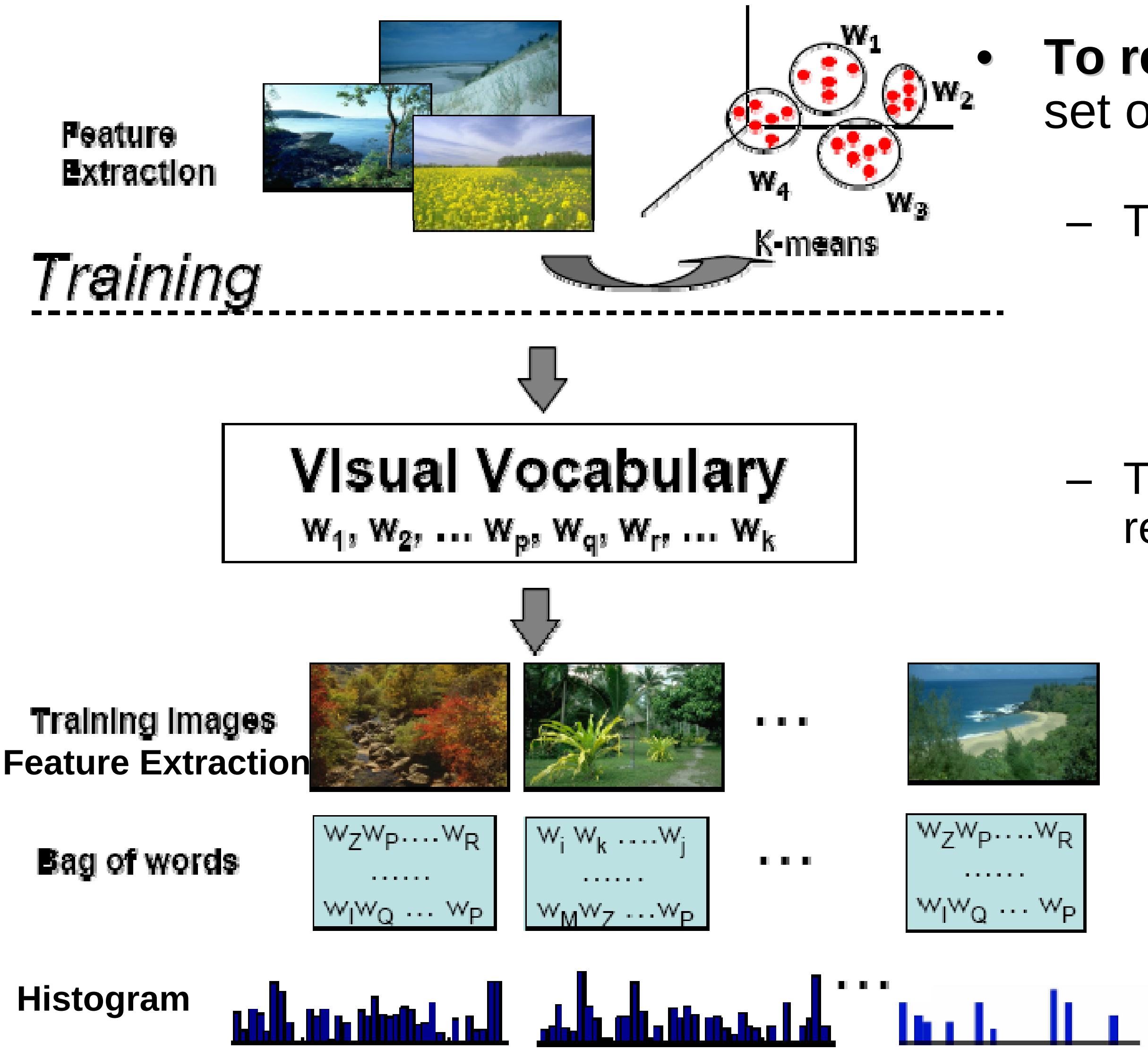
每个单词出现的次..
数(频率)

应用问题举例 2：图像检索

- 基于视觉词包模型建立图像的向量表示
  - – **Bag of Visual Words**



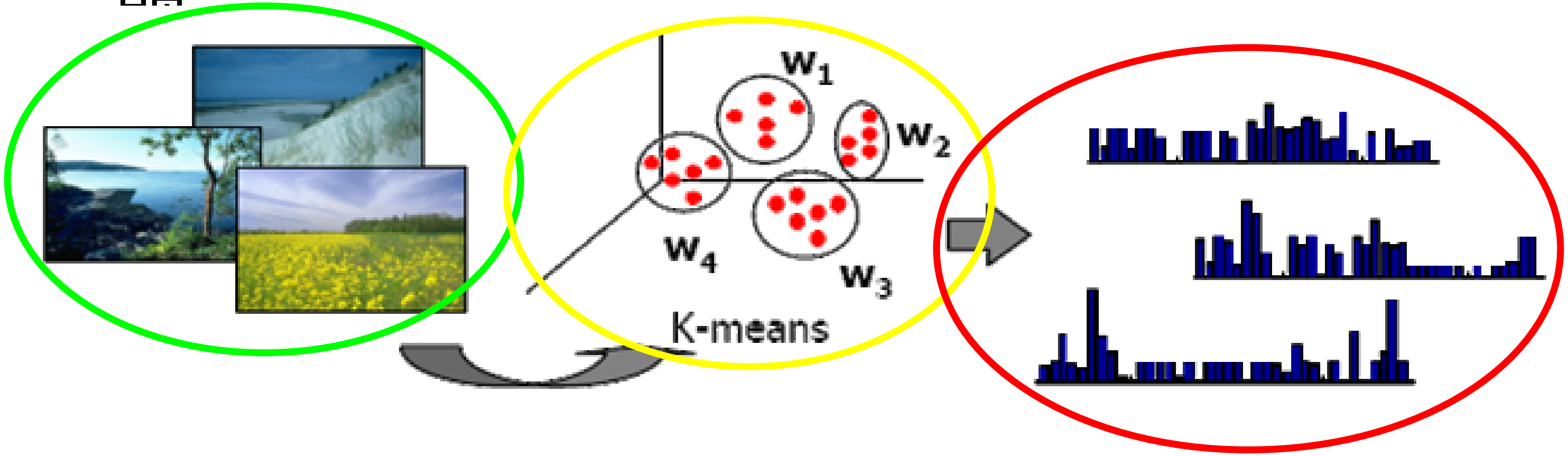
视觉词袋(Bag of Visual Words)


Feature Extraction
w_1
w_2
w_3
w_4
K-means
Training
Visual Vocabulary
w_1, w_2, … w_p, w_q, w_r, … w_k
Training Images Feature Extraction
…
Bag of words
$w_Z w_P \ldots w_R$ …… $w_I w_Q \ldots w_P$
$w_i\ w_k \ldots w_j$ …… $w_M w_Z \ldots w_P$
…
$w_Z w_P \ldots w_R$ …… $w_I w_Q \ldots w_P$
Histogram
…

- **To represent** an image via a set of ***local*** features
  - To extract local features
    - **Patches (i.e. block) sampling**
    - **Patches descriptor**
  - To construct a global representation
    - **Histogram**
      - **To construct codebook**
      - **To assign codeword**
      - **To construct distribution histogram**

应用问题举例 3：图像分类

- 一个复杂的应用问题往往需要一系列预处理步骤和特征抽取，也涉及到多种学习问题



  – VOC: 图像物体分类(Visual Objects Categorization)

- BoW: 词包(Bag of Words)

应用问题举例 4：动态场景分类

- 基于迁移的图像特征建立向量化表示

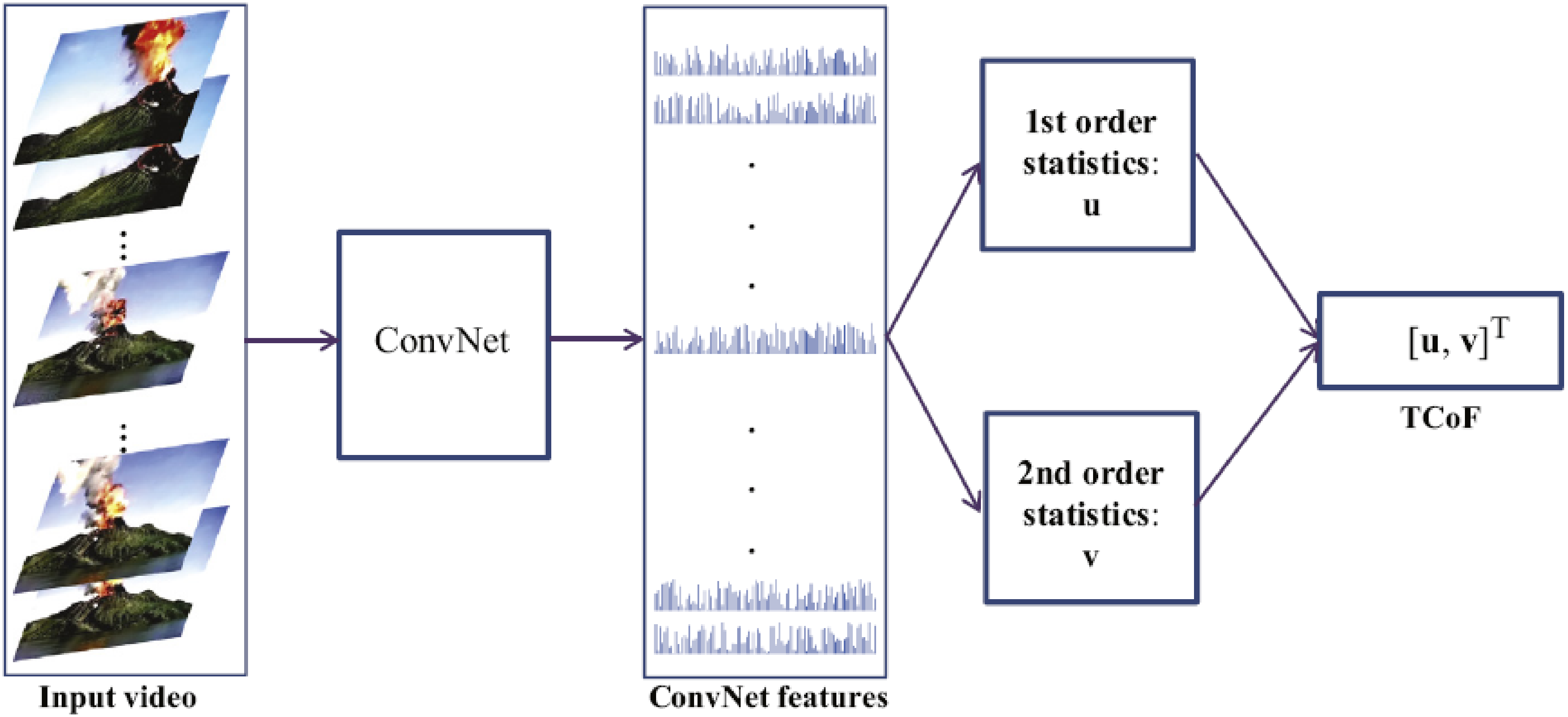
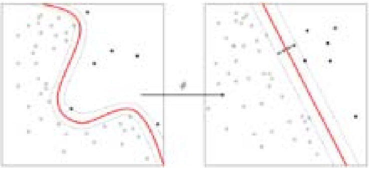
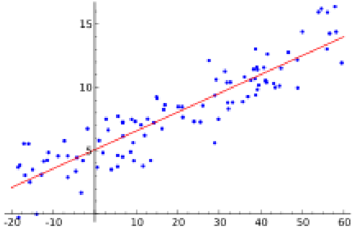


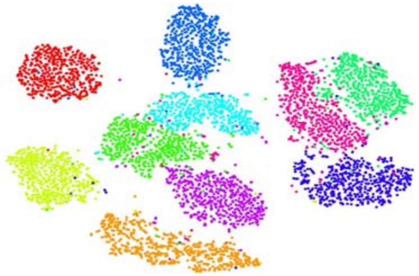
Fig. 4. An illustration of our TCoF scheme.

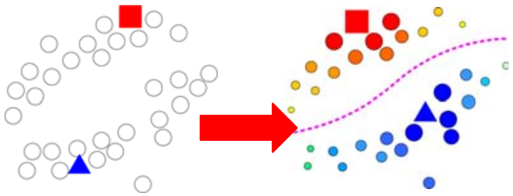
学习问题的基本形式





- 按照是否有监督信息使用
  - 有监督学习(Supervised learning)
    - 分类: 输出为离散变量
    - 回归: 输出为连续变量
  - 无监督学习(Unsupervised learning)
    - 聚类: 寻找数据的某种划分
    - 降维: 寻找数据的“坐标系”变换
  - 半监督学习(Semi-supervised learning)





- 表示符号说明
  - 训练数据(training data)

$$(X,Y)=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_n,y_n)\}$$

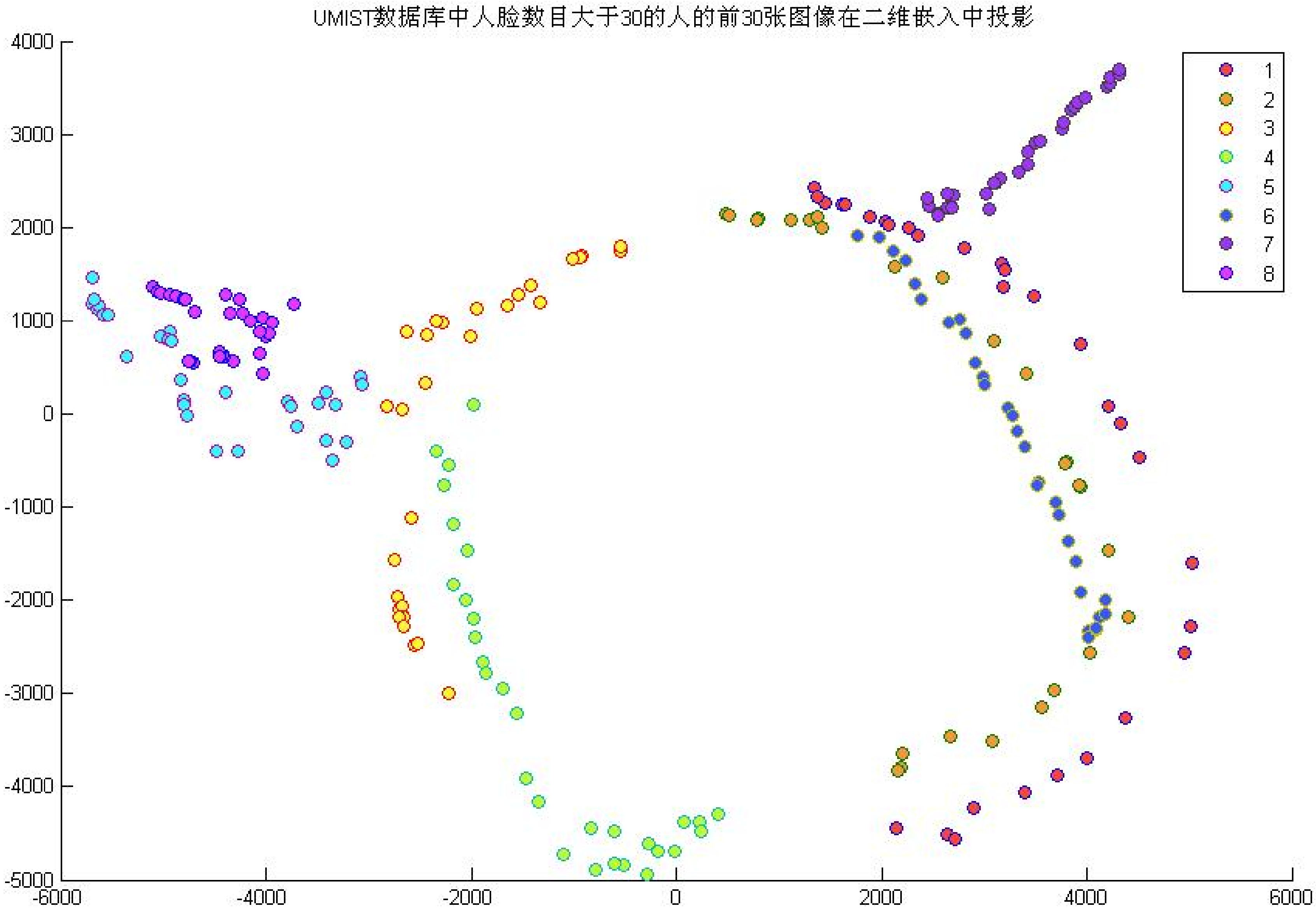
  - 数据本身

$$X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$$

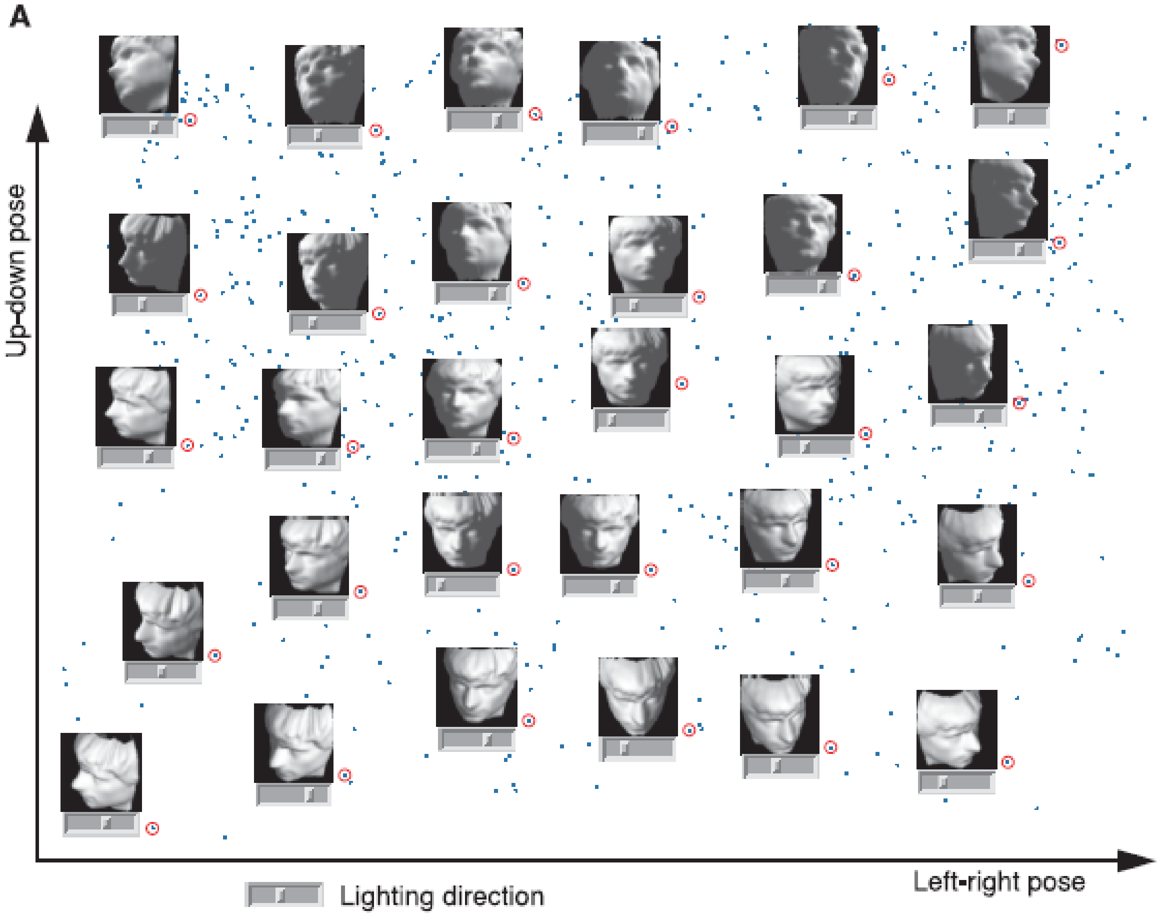
应用问题举例：II

- 数据降维
- 数据可视化
- 图像修复
- 背景与目标分离
- 歌声与背景音乐分离
- 视频中运动物体分割
- …

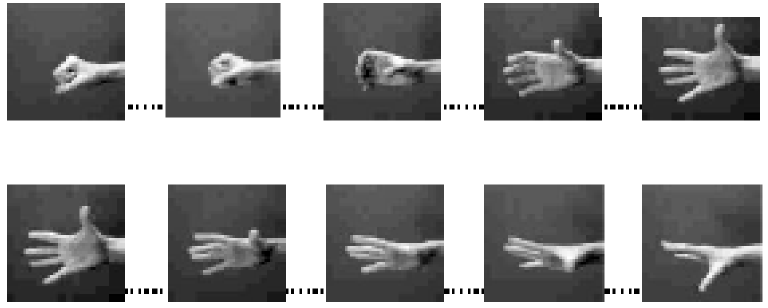
UMIST 数据集的可视化(by ISOMAP)



- 数据可视化



B



[1] J. B. Tenenbaum, V. de Silva, J. C. Langford: "A Global Geometric Framework for Nonlinear Dimensionality Reduction", Science, 2000.

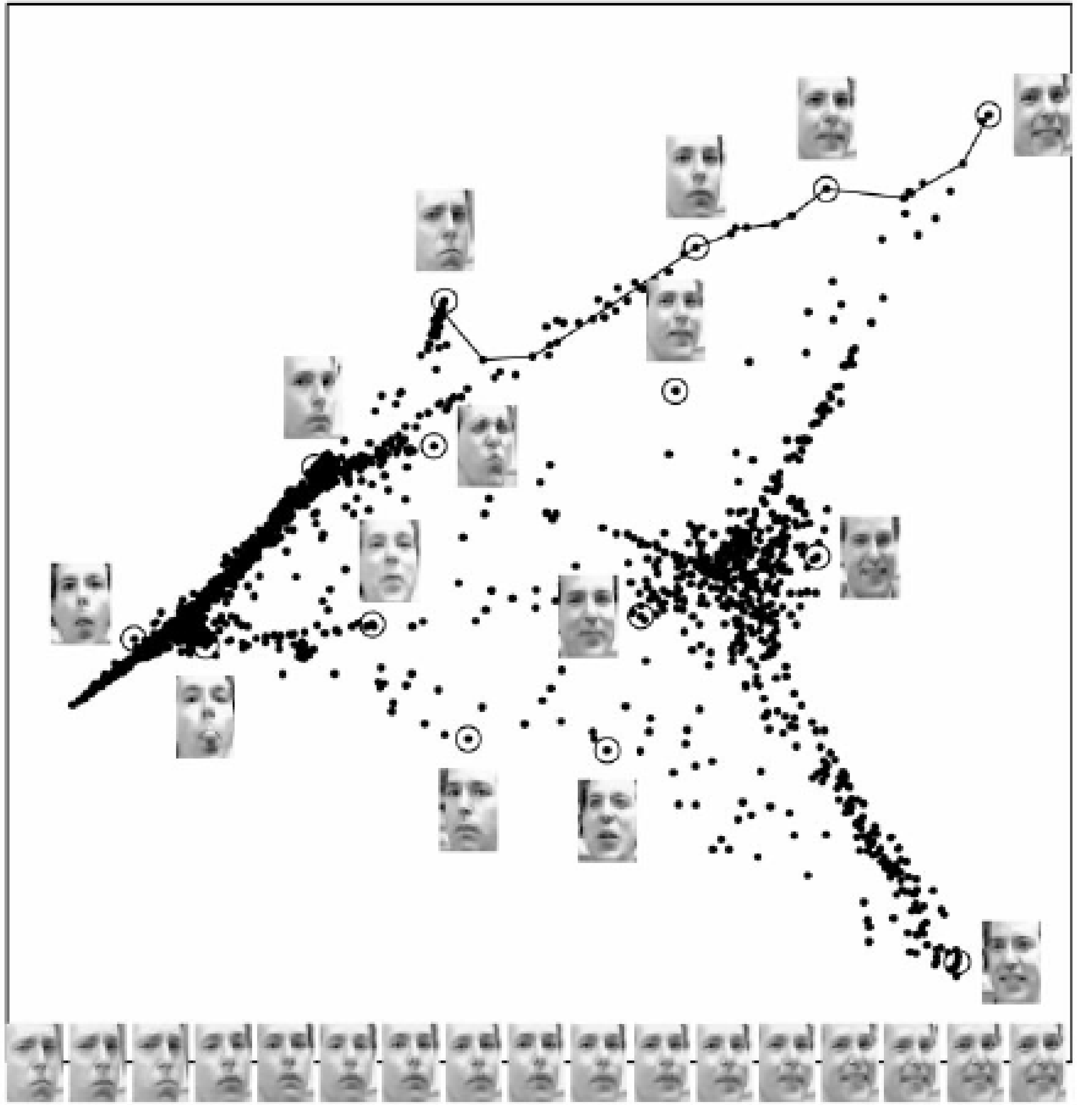
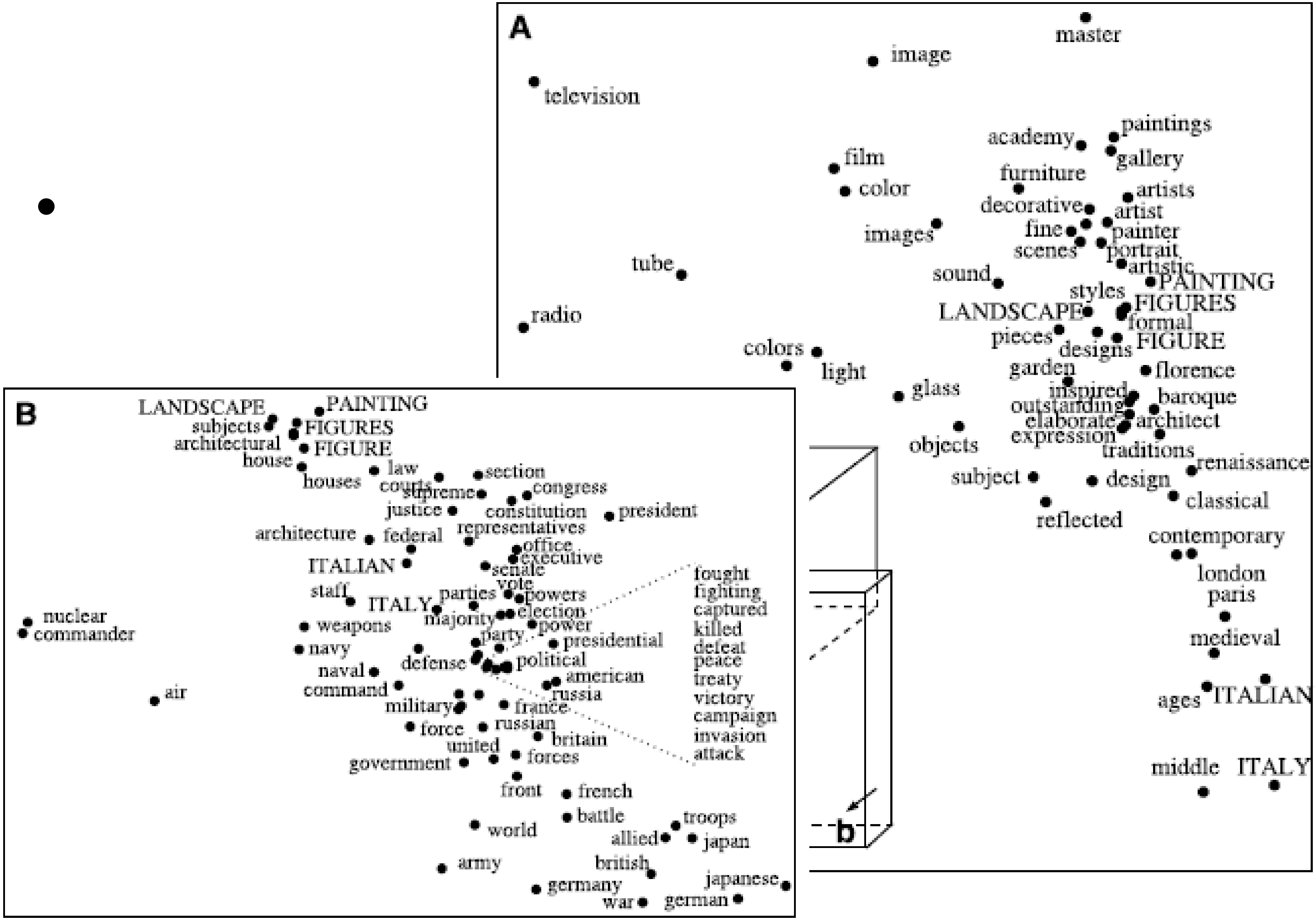


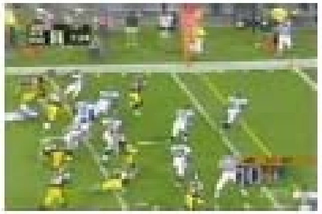
Fig. 3. Images of faces (11) mapped into the embedding space described by the first two coordinates of LLE. Representative faces are shown next to circled points in different parts of the space. The bottom images correspond to points along the top-right path (linked by solid line), illustrating one particular mode of variability in pose and expression.

[1] S. Roweis and L. Saul: "Nonlinear dimensionality reduction by locally linear embedding", Science, 2000.



[1] J. B. Tenenbaum, V. de Silva, J. C. Langford: "A Global Geometric Framework for Nonlinear Dimensionality Reduction", Science, 2000.

视频关键帧 (from TRECVID2006)



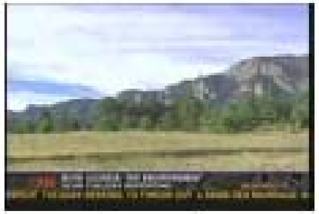
Sports

Weather

Office

Meeting

Desert



Mountain

Water

Corp. Leader

Police & Security

Military

Animal

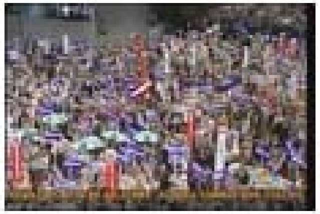
Screen

US Flag

Airplane

Car

Truck



People Marching

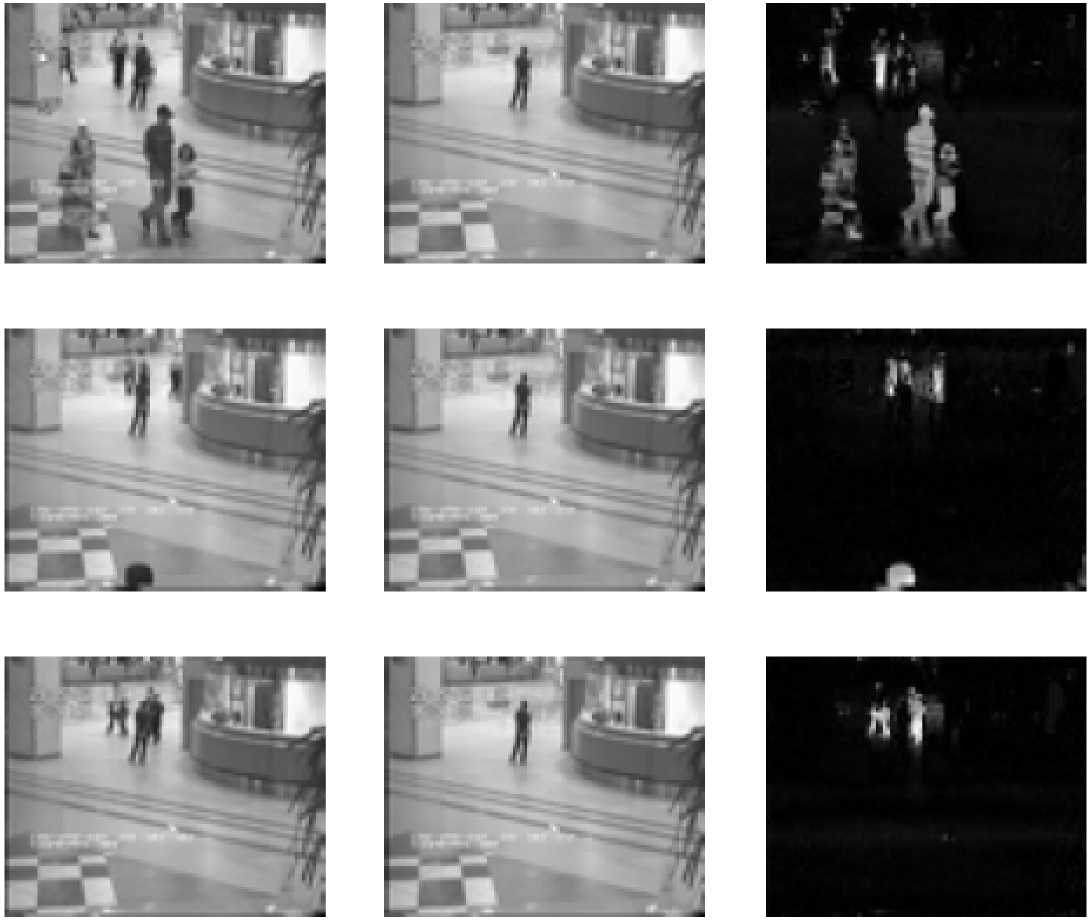
Explosion

Map

Charts

视频中背景与目标分离

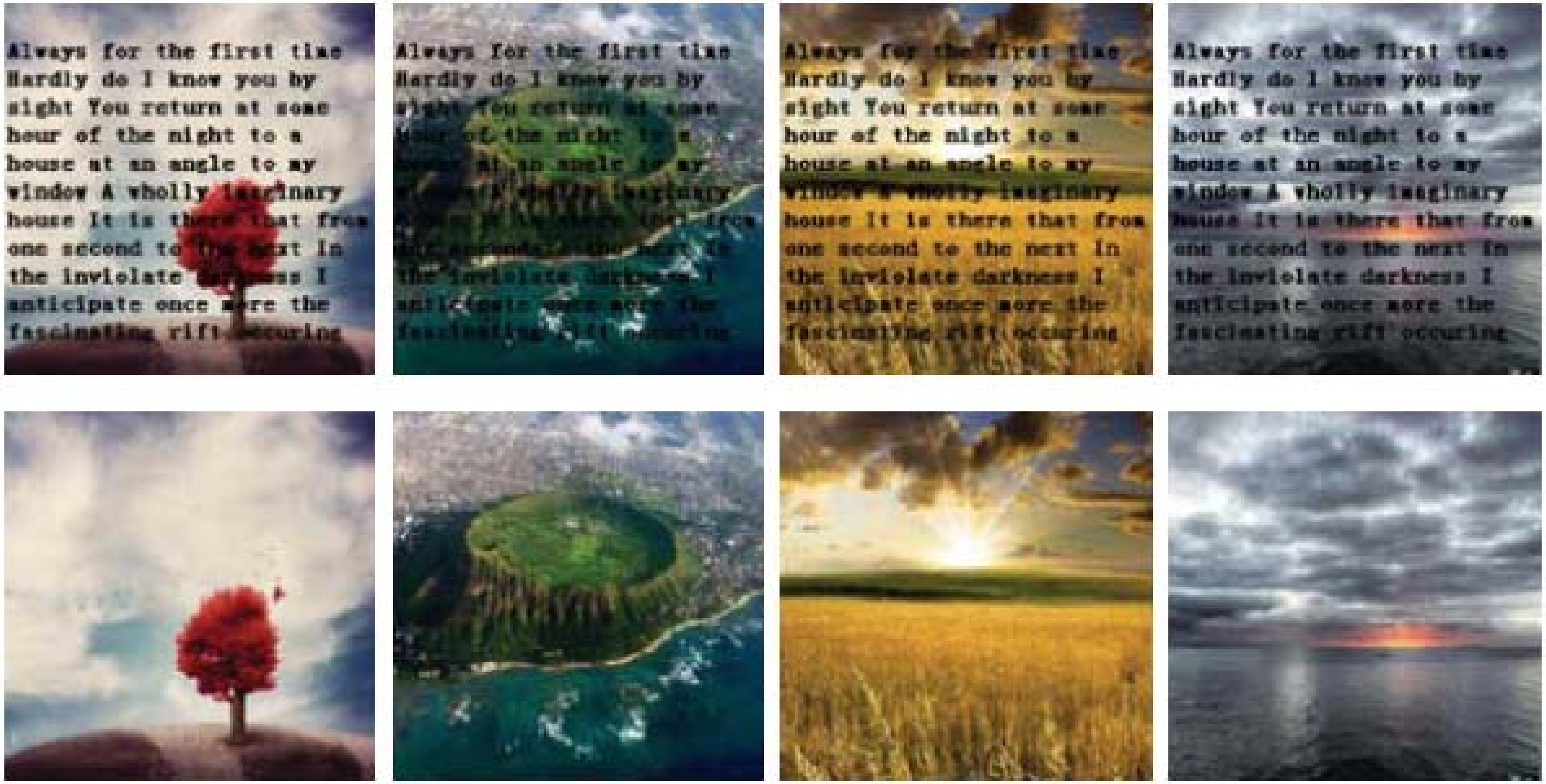
$$X \quad = \quad D \quad + \quad E$$



(a) (b) (c)

[1] J. Wright, A. Ganesh, S. Rao, Y. Peng, Y. Ma. "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization", NIPS, pp.2080-2088, 2009.

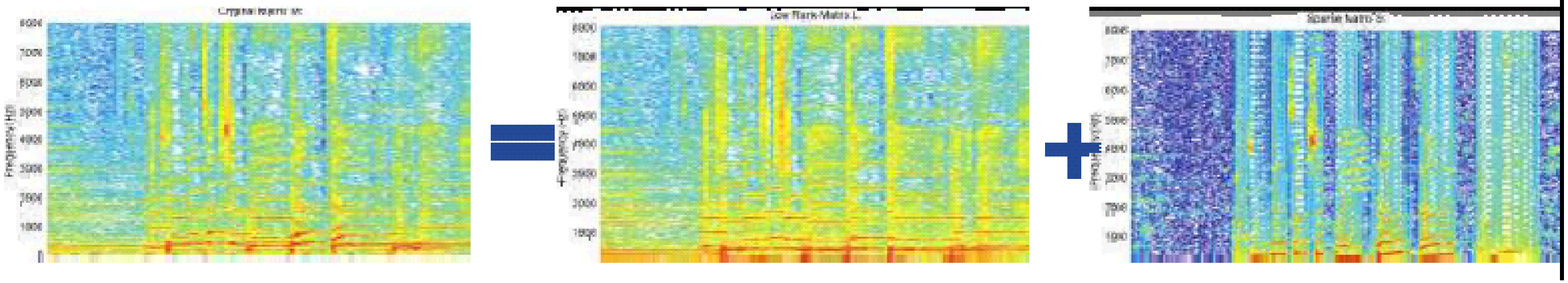
图像恢复(Image impainting)

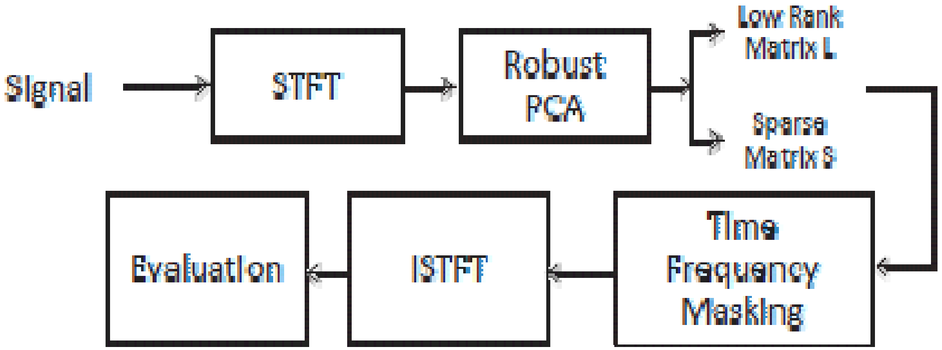
- 

Always for the first time Hardly do I know you by sight You return at some hour of the night to a house at an angle to my window A wholly imaginary house It is there that from one second to the next In the inviolate darkness I anticipate once more the fascinating rift occuring

歌声与背景音乐的分离

- Songs (STFT) Low-rank (music) Sparse (voices)





Signal

STFT

Robust PCA

Low Rank Matrix L

Sparse Matrix S

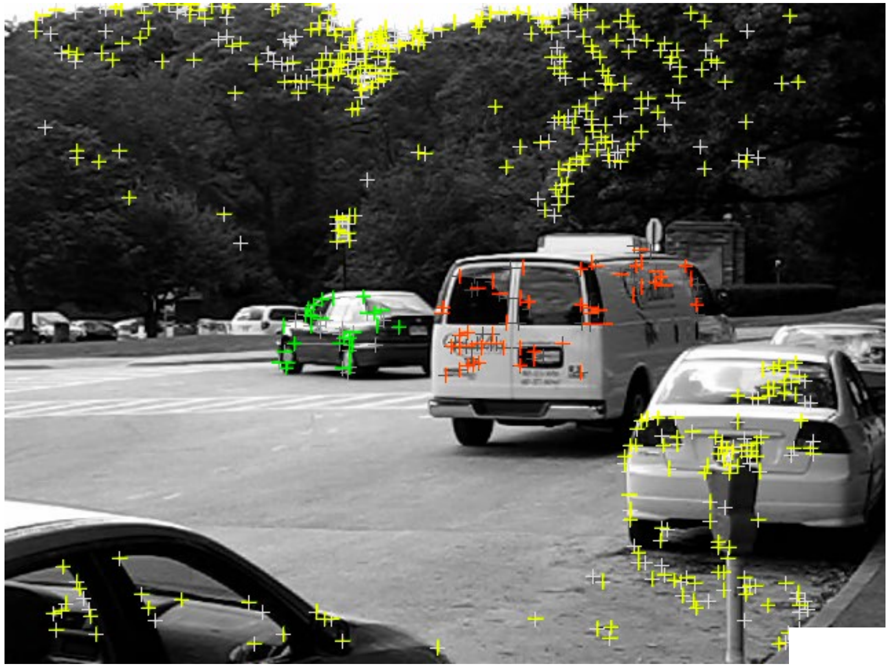
Time Frequency Masking

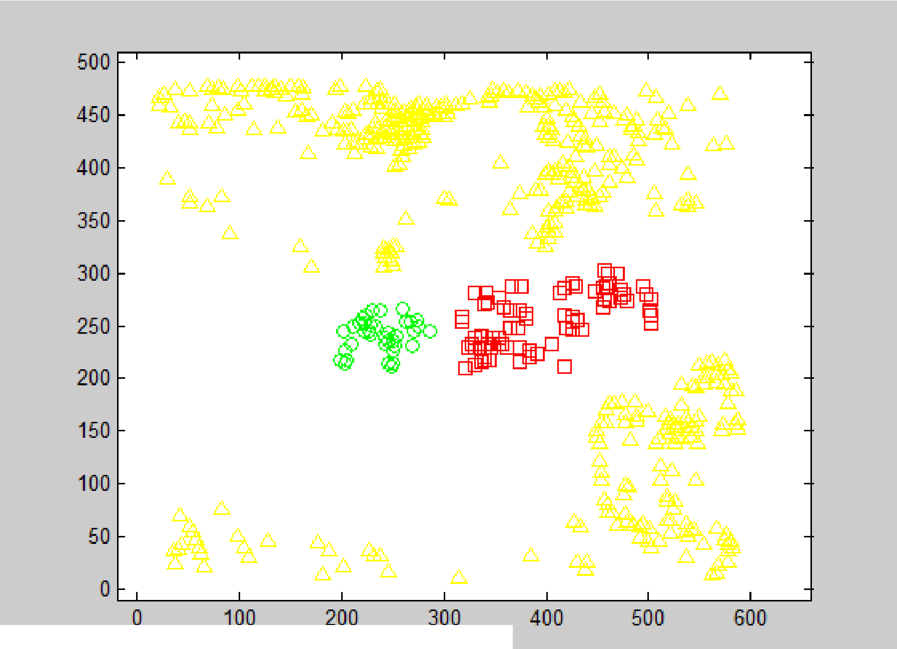
ISTFT

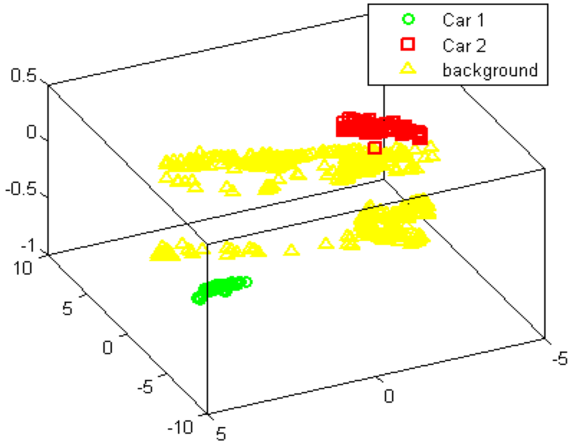
Evaluation

[1] Min, Zhang, Wright, Ma, CIKM 2010 / Sprechmann, Bronstein, Sapiro, ISMIR 2012

视频中的运动分割







[1] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," in IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

学习问题的其它形式

- 主动学习(Active learning)
  - 被动学习**(Passive learning)**
- 在线学习(Online learning)
  - 成批学习**(Batch learning)**
- 归纳学习(Inductive learning)
  - 传导学习**(Transductive learning)**
- 迁移学习(Transfer learning)
  - 借助一种学习去影响另一种学习
- 强化学习(Reinforcement learning)
  - 决策或控制领域，没有明确的监督信息，只能通过回报(reward)函数对每次决策作出反馈
- 深度学习(Deep learning)

Q / A

- Any Question? …

专题 一：基于实例的学习

- **内容提要**
  - 引言
  - 最近邻规则 **(Nearest Neighbor Rule)**
    - 非线性回归模型
  - 帕森窗**(Parzen Windows)**
    - 密度估计问题的引出
    - Kernels
    - 瓦森-纳达拉亚估计器(Watson-Nadaraya Estimator)
  - 应用问题举例**:**
    - MNIST数据集 / VOC 与 BoW模型
  - 大数定律？收敛速度**?**