时空行为理解

章毓晋

清华大学电子工程系 100084 北京

总目录

- 引言（+ 课程历史）
- 正文
  - 时空行为理解层次
  - 从时空兴趣点到轨迹
  - 动作/活动识别（示例）
  - 深度学习与行为识别（示例）
- 结语

引 言

- **图象：**

  用各种观测系统以不同形式和手段观测客观世界而获得的，可以直接或间接作用于人眼并进而产生视知觉的实体

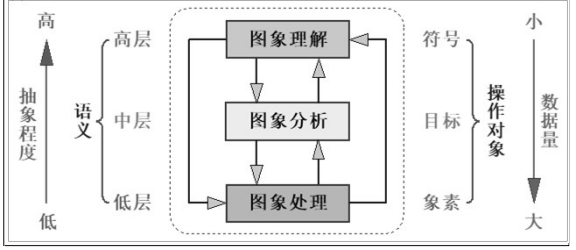
  图象（广义/抽象）⊃ 图像（狭义/具体）

- **图象和信息：**

  人类从外界（客观世界）获得的信息约有75%来自视觉系统

引 言

- **图象工程：**不同层次图象技术的有机结合及应用。 **三个层次**



引 言

- **图象工程文献综述系列**
  - 每年《中国图象图形学报》5月那一期
  - 已历时 24 年，涉及 15 种刊物，15095（62186）篇论文

主要目的

(1) 概括我国图象工程发展现状

(2) 帮助读者查阅有关研究文献

(3) 对期刊编者和论文作者提供参考

引 言

- **图象工程研究文献**
  - 图象处理（图象 ⇒ 图象）
  - 图象分析（图象 ⇒ 数据）
  - 图象理解（图象 ⇒ 解释）
  - 技术应用（图象处理、分析、理解技术的工程实现和应用）
  - 综述（综合图象处理、分析、理解）

引　言

- 图象处理
  - 图象获取
  - 图象重建（从投影重建）
  - 图象增强和恢复等
  - 图象（视频）压缩编码

从2000年起增加

- 图象数字水印和图象信息隐藏

引　言

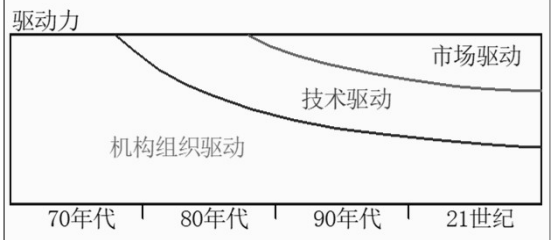
- 2000～2004（博士生学科前沿课）

国际标准MPEG-21
和
图象数字水印

- 2002年
  - ✓ 图象数字水印 ⇒ 图象工程教材

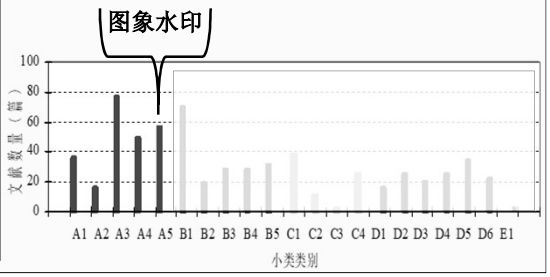
引　言

不同时代标准化的驱动力



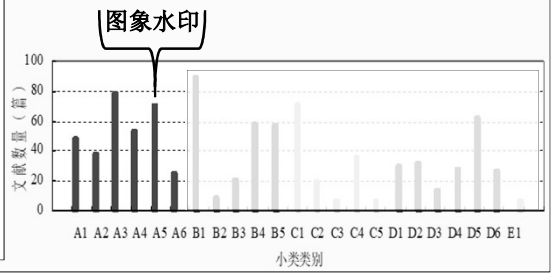
引　言

- 对2004年图象工程文献的统计



引　言

- 对2007年图象工程文献的统计



引　言

- 图象分析
  - 图象分割，边缘及角点等基元的检测
  - 目标表达、描述、测量
  - 目标特性的提取分析
  - 目标检测和识别

从2000年起增加

- 人体生物特征提取和验证（包括人体、人脸和器官的检测、定位与识别）

引　言

- 2005～2008（博士生学科前沿课）

人脸表情识别

- 2007年
  - ✓ 人脸和表情识别 ⇒ 图象工程教材 (第2版，中册)

引　言

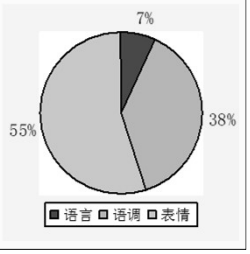
- 人脸表情分析

人类情感交流中的地位

- 语言内容：7%
- 语音语调：38%
- 脸部表情：55%

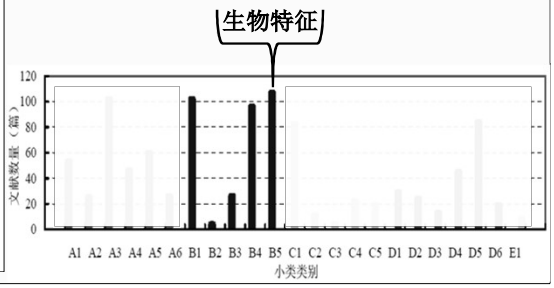
表情反映内心

- 测谎（>95%）



引　言

- 对2009年图象工程文献的统计



引　言

- 2009～2010
- （博士生学科前沿课）

基于子空间的人脸识别

程正东，贾彗星
李　乐，沈　斌
谭华春，严　严
章毓晋，朱云峰

引　言

第1章　绪论　　第2章　人脸检测
第3章　人脸跟踪　第4章　人脸描述
第5章　基本线性子空间方法
第6章　张量方法　第7章　核方法
第8章　非负矩阵（集）分解
第9章　分类器设计
第10章　评价指标与评测比较
附录A　张量　　附录B 3-D人脸识别综述
附录C　相关识别概述　附录D 常用数据库

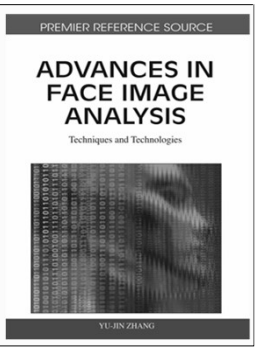
引　言

- 2011～2013
- （博士生学科前沿课）

人脸图象分析进展
—— 技术和应用

Yu-Jin ZHANG (Editor)

33 experts from
16 countries and regions



引　言

- 背景介绍　Introduction and Background
- 特征提取　Facial Feature Extraction
- 特征降维　Feature Dimensionality Reduction
- 人脸识别　Face Recognition
- 表情分类　Facial Expression Classification
- 不变技术　Invariance Techniques

引　言

➢ 图象理解

- 图象匹配和融合
- 场景恢复
- 图象感知和解释
- 基于内容的图象和视频检索

从2005年起增加

- 时空技术(高维运动分析、3-D 姿态检测、跟踪、举止判断和行为理解)

引　言

➢ 2014～2018（博士生学科前沿课）

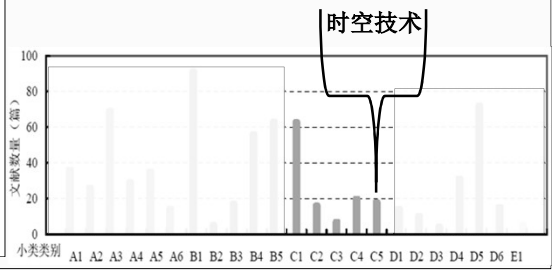
时空行为理解

➢ 2012年

✓ 时空行为理解 ⇒ 图象工程教材 (第3版，下册)

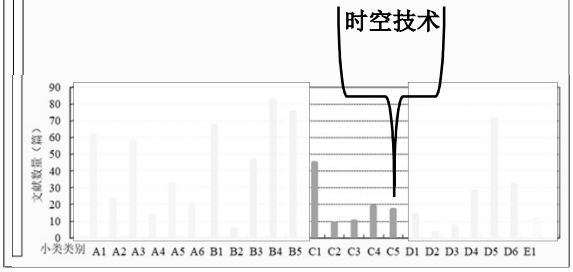
引　言

➢ 对2013年图象工程文献的统计



引　言

➢ 对2018年图象工程文献的统计



引　言

➢ 理解时空行为

✓ 时空：客观

- 位置、轨迹、速度、外观、姿态、关系、…
- 对群体目标活动中的聚合、消散、分化、合并等动态演变现象，…

✓ 行为：主观

- 举止、动向、态势、情感、…

© Y.J.ZHANG.

引　　言

计算机视觉前沿论坛（2013～）

徐光祐，刘允才，章毓晋。计算机视觉——探索行为理解，认知内心世界。中国图象图形学报，2013，18(2): 131

章毓晋。时空行为理解。中国图象图形学报，2013，18(2): 141-151（领跑者5000）

郑胤，陈权崎，章毓晋。深度学习及其在目标和行为识别中的新进展。中国图象图形学报，2014，19(2): 175-184 （领跑者5000）

© Y.J.ZHANG.

时空行为理解层次

从动作到行为的五个层次

(1) 动作基元（action primitives）

(2) 动作（action）

(3) 活动（activity）

(4) 事件（events）

(5) 行为（behavior）

指用来构建动作的原子单元，一般对应场景中局部短暂的运动信息

© Y.J.ZHANG.

时空行为理解层次

从动作到行为的五个层次

(1) 动作基元（action primitives）

(2) 动作（action）

(3) 活动（activity）

(4) 事件（events）

(5) 行为（behavior）

由主体的一系列动作基元构成的有具体意义的集合体（有序组合），一般动作常代表由一个人所进行的简单运动模式，且常仅持续秒的量级。人体动作的结果常导致人体姿态的改变

© Y.J.ZHANG.

时空行为理解层次

从动作到行为的五个层次

(1) 动作基元（action

(2) 动作（action）

(3) 活动（activity）

(4) 事件（events）

(5) 行为（behavior）

为完成某个工作或实现某个目标而由主体执行的一系列动作的组合（主要强调逻辑组合）。活动是相对大尺度的运动，一般依赖于环境和交互人。活动常代表由多个人参与的序列（可能交互的）复杂动作，且常持续较长的时段

© Y.J.ZHANG.

时空行为理解层次

从动作到行为的五个层次

(1) 动作基元（action primitives）

(2) 动作（action）

(3) 活动（activity）

(4) 事件（events）

(5) 行为（behavior）

指在特定时间段和特定空间位置发生的某种特定活动。通常其中的动作由多个主体/发起者执行（群体活动）。对特定事件的检测常与异常活动有关

© Y.J.ZHANG.

时空行为理解层次

从动作到行为的五个层次

(1) 动作基元（action primitives）

(2) 动作（action）

(3) 活动（activity）

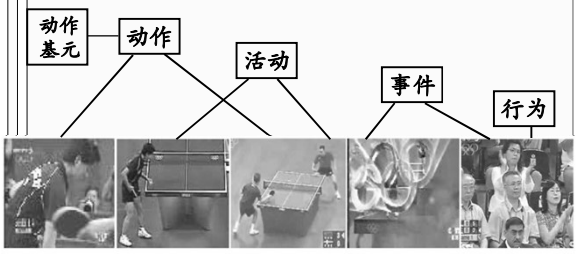
(4) 事件（events）

(5) 行为（behavior）

主体/发起者主要指人或动物，强调主体/发起者受思想支配而在特定环境/上下境中改变动作，持续活动和描述事件等

时空行为理解层次

五个层次（示例）



乒乓球比赛中的几个画面

从时空兴趣点到轨迹

1. 空间兴趣点的检测

使用线性尺度空间表达对图象建模

$$L^{sp}(x,y;\sigma_l^2)=g^{sp}(x,y;\sigma_l^2)\otimes f^{sp}(x,y)$$

高斯核

$$g^{sp}(x,y;\sigma_l^2)=\frac{1}{2\pi\sigma_l^2}\exp[-(x^2+y^2)/2\sigma_l^2)]$$

Harris兴趣点检测

图象点邻域中朝向分布矩阵

$$\mu^{sp}(\bullet;\sigma_l^2,\sigma_l^2)=g^{sp}(\bullet;\sigma_l^2)\otimes\{[\nabla L(\bullet;\sigma_l^2)][\nabla L(\bullet;\sigma_l^2)]^T\}$$

$$=g^{sp}(\bullet;\sigma_l^2)\otimes\begin{bmatrix}(L_x^{sp})^2 & L_x^{sp}L_y^{sp}\\ L_x^{sp}L_y^{sp} & (L_y^{sp})^2\end{bmatrix}$$

从时空兴趣点到轨迹

1. 空间兴趣点的检测

μ^{sp}的本征值λ_1和λ_2（$\lambda_1 \le \lambda_2$）构成f^{sp}沿两个图象坐标方向而变化的描述符

检测角点函数的正极大值

$$H^{sp}=\det(\mu^{sp})-k\bullet\mathrm{trace}^2(\mu^{sp})=\lambda_1\lambda_2-k(\lambda_1+\lambda_2)^2$$

在感兴趣点，本征值的比$a=\lambda_2/\lambda_1$应该很大

对H^{sp}的正局部极值，a应该满足$k \le a/(1+a)^2$

设$k = 0.25$，H的正最大值将对应理想的各向同性兴趣点（此时$a=1$，即$\lambda_1=\lambda_2$），k小 $\Rightarrow$ a大

从时空兴趣点到轨迹

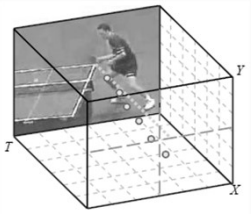
2. 时空兴趣点的检测

检测在局部时空体中具有沿时和空都有图象值较大变化的位置

$$L(\bullet;\sigma_l^2,\tau_l^2)=g(\bullet;\sigma_l^2,\tau_l^2)\otimes f(\bullet)$$

空间方差σ_l^2和时间方差τ_l^2

为了检测感兴趣点，在f中搜索具有μ的显著本征值λ_1，λ_2，λ_3的区域



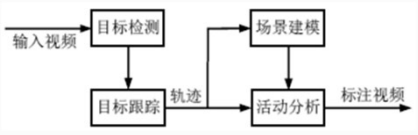
时空兴趣点示例

从时空兴趣点到轨迹

通过对（监控）场景中各个运动目标行为的描述和刻画来提供对场景状态的把握

动态轨迹学习和分析：

先对感兴趣目标进行检测并跟踪，接着用所获得的轨迹自动地构建场景模型，最后用该模型描述监控的状况并提供对活动的标注

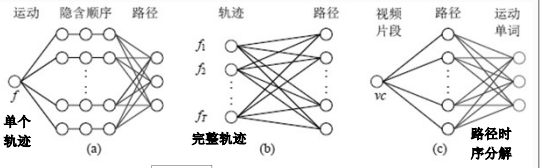


从时空兴趣点到轨迹

3. 活动路径学习

活动定义在开始和结束的两个感兴趣点之间

3种基本结构的主要区别包括输入的种类，运动矢量，轨迹/视频片段，以及运动抽象的方式



轨迹和路径学习方案

© Y.J.ZHANG.

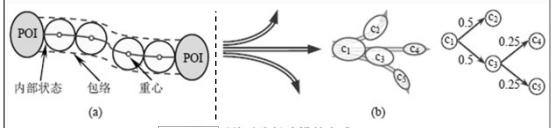
从时空兴趣点到轨迹

4. 活动路径建模

路径模型是对聚类的紧凑表达（图模型推理）

① 考虑完整的路径，有平均的中心线，两边还有包络指示路径范围

② 将路径分解为子路径（表示成子路径的树）

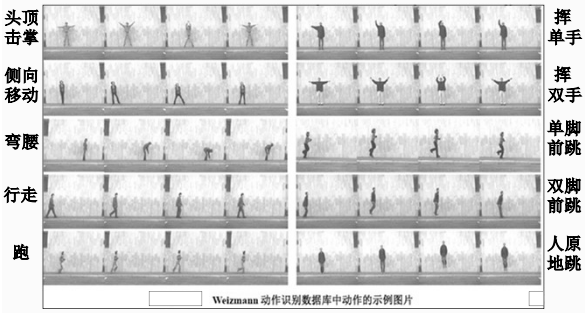


两种对路径建模的方式

© Y.J.ZHANG.

动作/活动识别（示例）

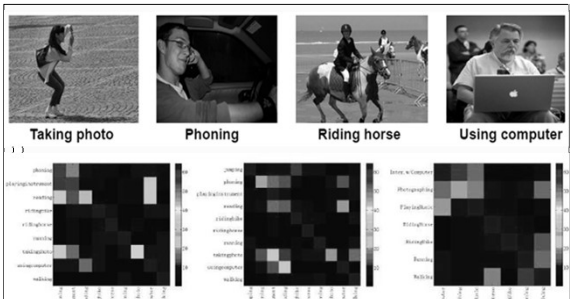
动作识别数据库



Weizmann 动作识别数据库中动作的示例图片

© Y.J.ZHANG.

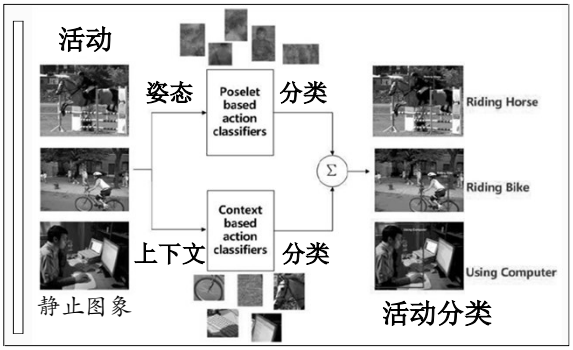
动作/活动识别（示例）



Zheng Y, Zhang Y-J, Li X, Liu B D. "Action Recognition in Still Images Using a Combination of Human Pose and Context Information". Proc. ICIP, 785-788, 2012

© Y.J.ZHANG.

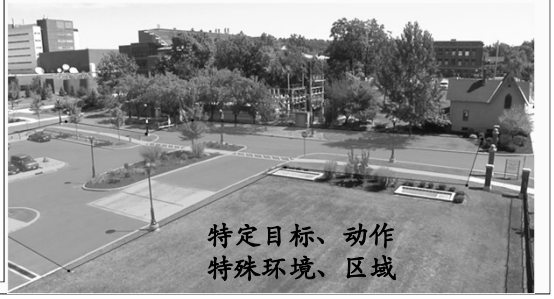
动作/活动识别（示例）



© Y.J.ZHANG.

动作/活动识别（示例）

停车场的人车互动



特定目标、动作
特殊环境、区域

© Y.J.ZHANG.

动作/活动识别（示例）

停车场的人车互动

人上车、人下车；车与人的相互位置关系

感兴趣的区域

动作/活动识别（示例）

停车场的人车互动

车的检测

（位置，状态，环境，相互关系，……）

车遮挡重叠？

车外观相似？

动作/活动识别（示例）

停车场的人车互动 人的检测和定位

动作/活动识别（示例）

停车场的人车互动 类别和描述

Table 1. Generated events. In event signatures, "h" stands for a human ID, and "v" for a vehicle ID; "f" is the frame number at which the event occurs.

Event signature	Description
Appears(h,f)*	A human appears inside the ROI
Disappears(h,f)*	A human disappears inside the ROI
IsNear(h,v,f)**	The human is near the vehicle
IsStopped(v,f)**	The vehicle is currently stopped

*: Not generated when the object intersects the ROI boundary.
**: Generated only if "Appears" or "Disappears" event is generated at the same frame.

动作/活动识别（示例）

停车场的人车互动 规则和效果

1. $Appears(h,f) \wedge IsStopped(v,f) \wedge IsNear(h,v,f) \Rightarrow GetsOut(h,v,f)$
2. $Disappears(h,f) \wedge IsStopped(v,f) \wedge IsNear(h,v,f) \Rightarrow GetsIn(h,v,f)$
3. $GetsIn(h,v1,f) \Leftrightarrow \neg GetsIn(h,v2,f)$
4. $GetsOut(h,v1,f) \Leftrightarrow \neg GetsOut(h,v2,f)$

Table 3. Inference results on VIRAT dataset

Event	Precision	Recall	F-score
GetsIn	31,58%	85,71%	46,15%
GetsOut	30,00%	75,00%	42,86%

Zhang Y-J, Vincent T. "Spatial-Temporal Behaviour Understanding". Proceedings of the 2nd International Conference on Electronics, Communications and Control, 2012

动作/活动识别（示例）

监控中的骑抢检测

监控摄象头遍布各处：

公共汽车、超市、车载、地铁、酒店、小区、路口、商场、停车场、银行

动作/活动识别（示例）

监控中的骑抢检测

袁静，章毓晋．融合梯度差信息的稀疏去噪自编码网络在异常行为检测中的应用．自动化学报，2017，43(4)：604-610

© Y.J.ZHANG.

动作/活动识别（示例）

监控中的骑抢检测　　视频序列

© Y.J.ZHANG.

动作/活动识别（示例）

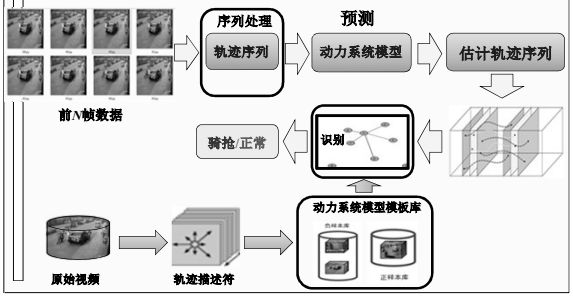
监控中的骑抢检测

定位和跟踪（人、车、背景）

© Y.J.ZHANG.

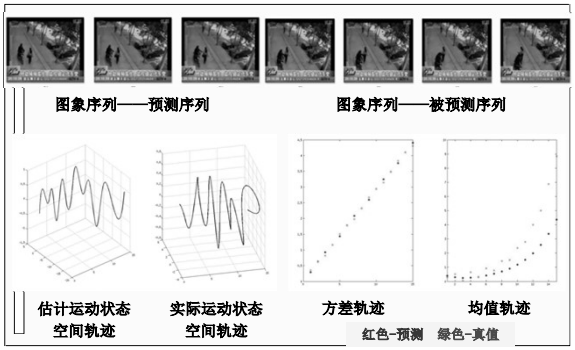
动作/活动识别（示例）

监控中的骑抢检测　　流程和步骤



© Y.J.ZHANG.

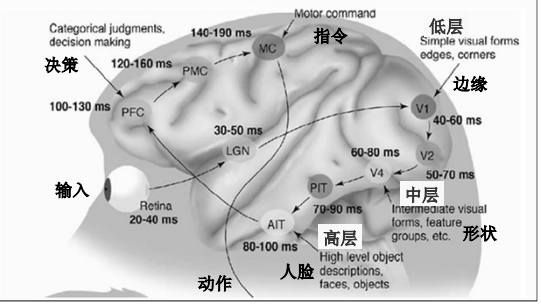
动作/活动识别（示例）



© Y.J.ZHANG.

深度学习与行为识别

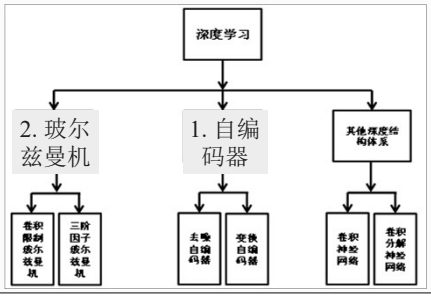
人脑分层次认知



© Y.J.ZHANG.

深度学习与行为识别

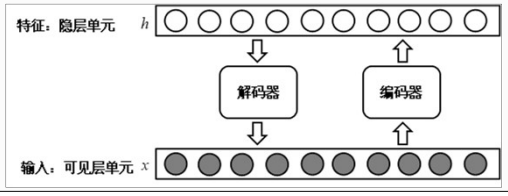
深度学习算法体系结构



深度学习与行为识别

深度学习算法体系结构

1、自编码器的核心思想是将输入信号进行编码，使用编码之后的信号重建原始信号，目的是让重建信号与原始信号相比重建误差最小



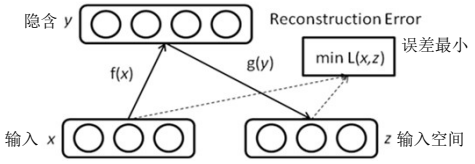
深度学习与行为识别

- The Classical Autoencoder
  - mapping visible input x to hidden representation y

$$y = f(x) = s(Wx + b)$$

  - mapping hidden representation y back to input space

$$z = g(y) = s(W'y + b')$$

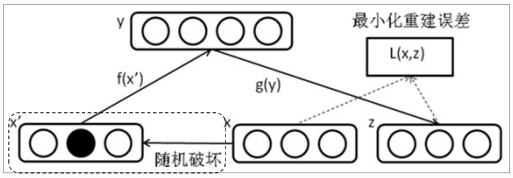


深度学习与行为识别

深度学习算法体系结构

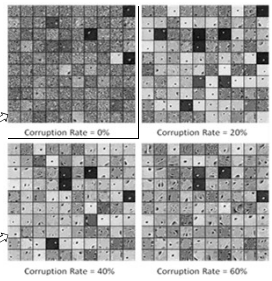
去噪自编码器（denoising autoencoder）

能发现输入中更多隐含的特征，能减小模型对输入数据的依赖性，能增强模型的鲁棒性



深度学习与行为识别

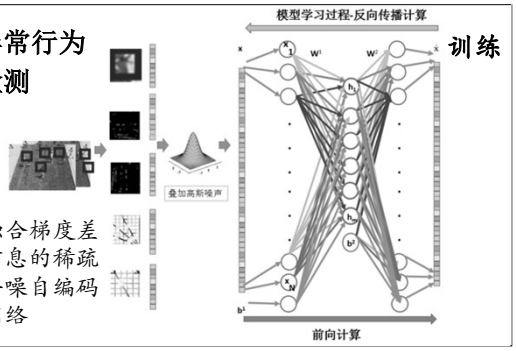
- Comparison between Classical Autoencoder and Denoising Autoencoder
  - Four examples with different corruption rate
  - Classical Autoencoder: the bases seem to be influenced by some spatial noise
  - Denoising Autoencoder: smoother and clearer



You Q H Z, Zhang Y-J. "A New Training Principle for Stacked Denoising Autoencoders". Proceedings of the 7th International Conference on Image and Graphics, pp.384-389, 2013.

深度学习与行为识别

异常行为检测

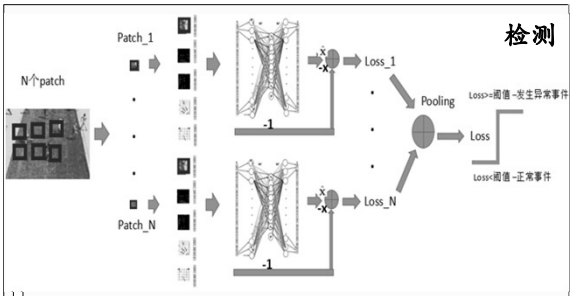


训练

融合梯度差信息的稀疏去噪自编码网络

深度学习与行为识别

检测



袁静，章毓晋。融合梯度差信息的稀疏去噪自编码网络在异常行为检测中的应用。自动化学报，2017，43(4)：604-610。

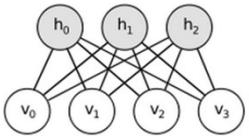
深度学习与行为识别

深度学习算法体系结构

2、限制/约束玻尔兹曼机是构成深度置信网络的基础单元，其本质是使得学习到的模型产生符合条件样本的概率最大

可以对全连通的玻尔兹曼机进行简化，使同层单元彼此独立（约束）

层内单元之间没有连接关系，层间单元是全连接关系

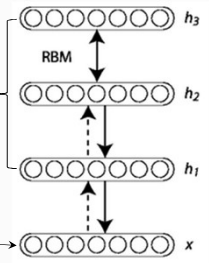


深度学习与行为识别

深度置信网络（Deep Belief Networks）

将约束玻尔兹曼机模型堆叠起来可构成一个深层置信网络，这样的几何模型可以用来提取训练数据中深层结构的特征

深层置信网络模型描述了可观测向量和所有隐藏层之间的联合概率分布



深度学习与行为识别

深度网络用于目标和行为识别

谷歌的虚拟人脑

使用 1000 台电脑构造出了包含 10 亿个连接的深达9层的自编码器“神经网络”

具有一定的自动学习能力

可以模拟一些人脑的功能

在完全没有标签的情况下，该网络能够自动地从训练集中学习到某些高层概念（如去“认识”动物“猫”，将猫的图片与cat联系起来）

深度学习与行为识别

深度网络用于目标和行为识别

大规模目标识别

在Image Net数据库（超过10,000,000张图象，超过1000个类别）上进行的大规模视觉识别比赛

早期最有效的算法

基于词袋模型（bag of words）

可变部件模型（deformable part model）

2012年基于深度学习的模型取得了最好的效果（超出了第二名至少10个百分点）

深度学习与行为识别

深度网络用于目标和行为识别

图象的同时分类和标注

图象分类：指的是对图象内容做整体的描述，例如给定一幅图象确定它属于“海滩”、“厨房”、“卧室”等预先定义好的类别中的哪一类

图象标注：指的是对于图象中包含的内容作出判断，例如一幅图象中是否包含“天空”、“汽车”、“树木”等预先定义好的目标，“骑车”，“篮球比赛”等预先定义好的活动，……

深度学习与行为识别

深度网络用于目标和行为识别

图象的同时分类和标注

Zheng Y, Zhang Y-J, Larochelle H. “Topic Modeling of Multimodal Data: An Autoregressive Approach”. Proceedings of Computer Vision and Pattern Recognition, 1370-1377, 2014

Zheng Y, Zhang Y-J, Larochelle H. “A Deep and Autoregressive Approach for Topic Modeling of Multimodal Data”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(6): 1056-1069, 2016

深度学习与行为识别

深度网络用于目标和行为识别

图象的同时分类和标注

使用基于神经自回归分布估计器（NADE）的监督性神经自回归分布主题模型（SupDocNADE）来处理同时图象分类和标注问题

NADE (**N**eural **A**utoregressive **D**istribution **E**stimator)

DocNADE (**Doc**ument)

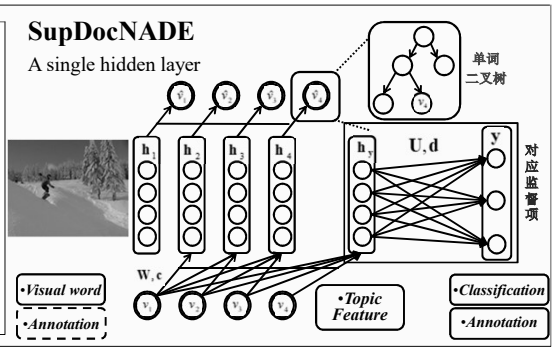
SupDocNADE (**Sup**ervised)

SupDeepDocNADE (**Deep**)

深度学习与行为识别

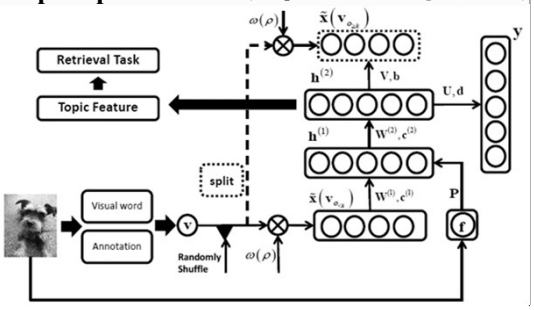
SupDocNADE

A single hidden layer

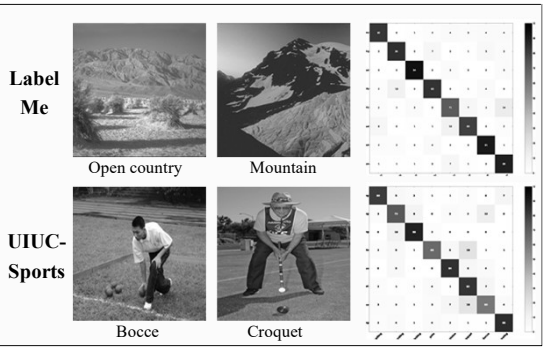


深度学习与行为识别

SupDeepDocNADE (Deep extension of SupDocNADE)



深度学习与行为识别



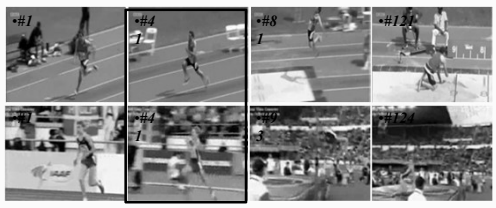
深度学习与行为识别

Flicker子集（2.5万张图象, 38 类）上的结果

Model	MAP
TF-IDF	0.384 ± 0.004
Multiple Kernel Learning SVMs [6]	0.623
TagProp [31]	0.640
Multimodal DBM [13]	0.651 ± 0.005
MDRNN [15]	0.686 ± 0.003
SupDeepDocNADE (1 hidden layer, 625 epochs pretraining)	0.654 ± 0.004
SupDeepDocNADE (2 hidden layers, 625 epochs pretraining)	0.671 ± 0.006
SupDeepDocNADE (3 hidden layers, 625 epochs pretraining)	0.670 ± 0.005
SupDeepDocNADE (2 hidden layers, 2325 epochs pretraining)	0.682 ± 0.005
SupDeepDocNADE (3 hidden layers, 2325 epochs pretraining)	0.686 ± 0.005
SupDeepDocNADE (2 hidden layers, 4125 epochs pretraining)	0.684 ± 0.005
SupDeepDocNADE (3 hidden layers, 4125 epochs pretraining)	$\mathbf{0.691} \pm 0.005$

深度学习与行为识别

视频中的人体活动分类 **问题和动机**



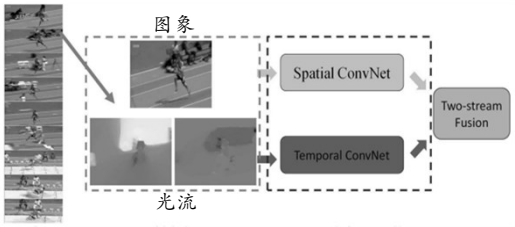
在助跑阶段很难区分，需要考虑整个视频段的顺序结构

Chen Q Q, Zhang Y-J. "Sequential Segment Networks for Action Recognition". IEEE Signal Processing Letters, 24(5): 712-716, 2017

深度学习与行为识别

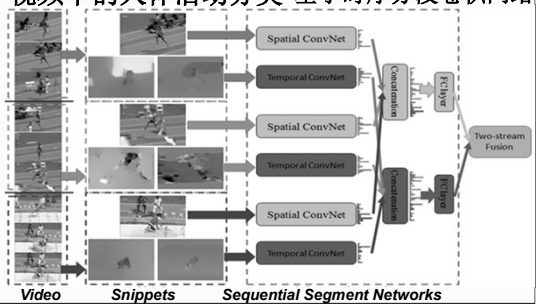
视频中的人体活动分类 基于双通道卷积网络

每次识别一个片段，将各片段平均作为最终结果



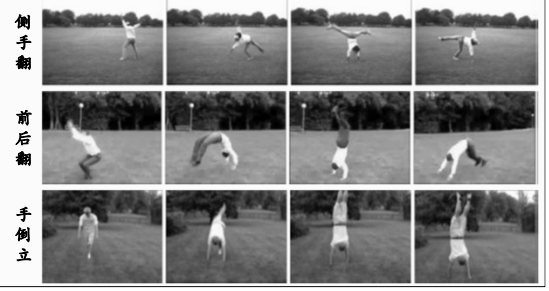
深度学习与行为识别

视频中的人体活动分类 基于时序分段卷积网络



深度学习与行为识别

视频中的人体活动分类 HMDB51库相似矩阵

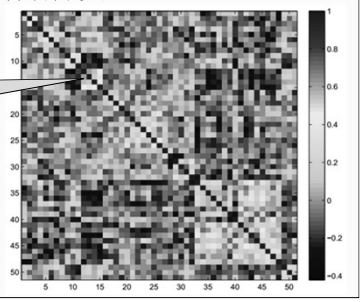


深度学习与行为识别

视频中的人体活动分类 HMDB51库相似矩阵

包含了人体翻转的活动

通过学习获得了时间结构上的相似信息



结　　语

- 引入一个有前途的研究方向
  - ✓ 对应高层的图象理解
- 介绍了一些相关研究和进展
  - ✓ 从客观到主观与客观结合
  - ✓ 从基元到行为（五个层次）
  - ✓ 从点到线到面到体（轨迹、路径、 …）
  - ✓ 从数学物理到（结合）神经科学
- 未来发展趋势

联 系 信 息

- 通信地址：北京清华大学电子工程系（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
- 电子邮箱：zhang-yj@tsinghua.edu.cn
- 个人主页：oa.ee.tsinghua.edu.cn/~zhangyujin/