CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

模式识别国家重点实验室
National Lab of Pattern Recognition

中国科学院自动化研究所
Institute of Automation Chinese Academy of Sciences

面向远距离人群感知的视频分析

王 亮

中国科学院自动化研究所

2017年4月22日　厦门

背 景

- 公共安全、国家安全、交通运输等诸多国计民生领域急迫需求**远距离人群感知**共性技术

我国社会深刻转型，公共安全事件高发、挑战严峻，迫切需要相关视频分析技术的有力支撑

远距离人群感知

通过对视频进行深度分析与理解，全面掌控远距离人群的身份、属性、行为、事件等信息

挑 战

距离远 ⇨ 看不清

描述人群表观、运动信息的有效像素很少，无法直接适用于传统算法

人数多 ⇨ 数不清

人群密度大、分布不均衡、遮挡严重等因素给群体分析、密度估计等任务带来困难

差异大 ⇨ 认不准

在任意场景、光照、姿态、表情、动作等非受控环境下，同一个体表观差异较大

范围大 ⇨ 找不准

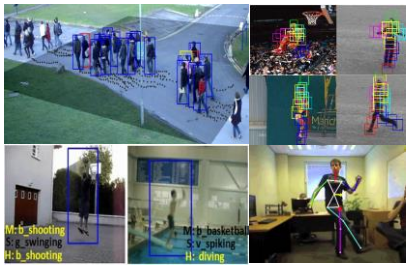
在海量目标、复杂背景等影响下，准确快速地查找特定个体、人群变得极为困难

任务分解

视频超分辨率、去噪、去模糊等

视频中人脸检测、行人检测、分割等

非约束环境下行人识别、属性、检索等



大范围目标跟踪、运动估计、行为识别等

底层视觉质量增强

个体检测与分割

个体认证与检索

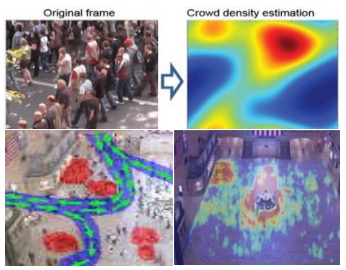
跟踪与行为识别

高层语义交互

群体密度流向估计

群体再识别

群体行为事件预测



复杂场景群体密度估计、流向预测等

大范围群体再识别、多群体关联分析等

非结构化群体行为分析、事件预警等

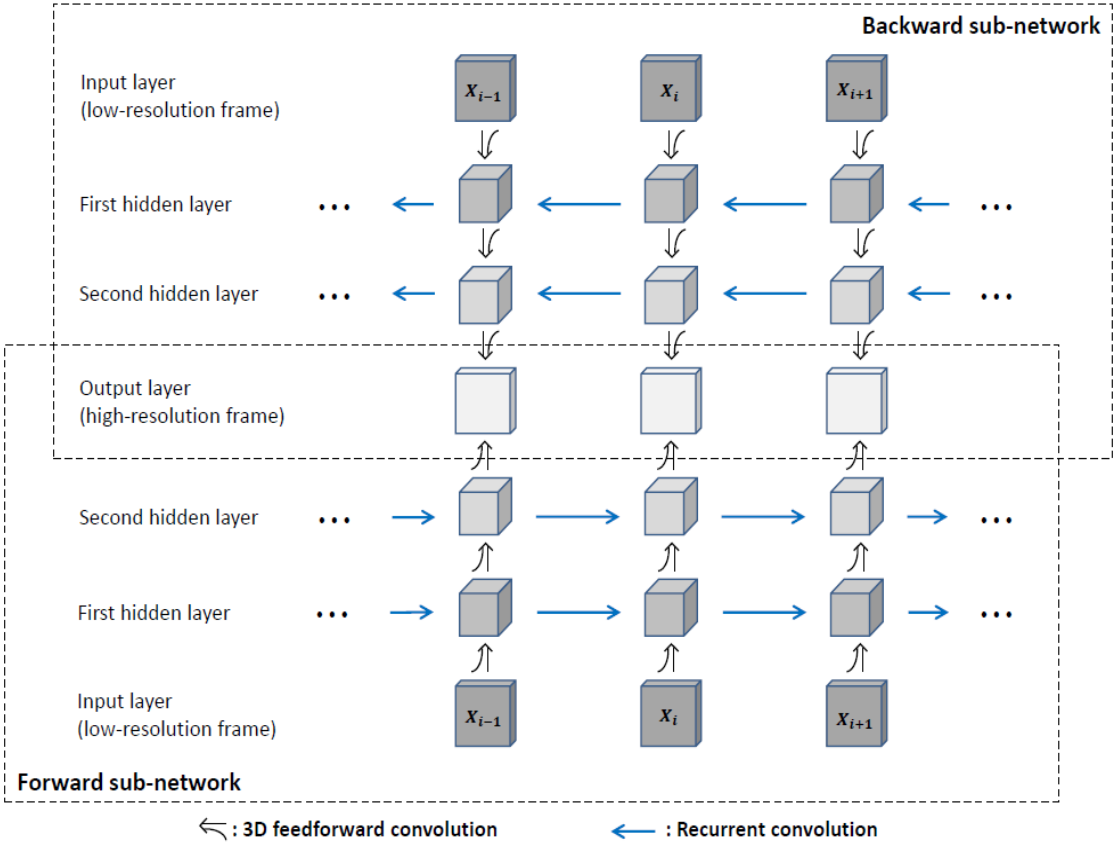
基于语义属性的视频问答、视频描述生成等

视频质量增强

基于双向循环卷积网络的视频超分辨率

任务：将远距离采集的低分辨率视频恢复到高分辨率视频

方法：提出全卷积双向循环网络来对视频中时间相依关系进行建模



: 3D feedforward convolution

对同一时刻下高低分辨率视频帧之间的空间结构相关关系进行建模

对邻近时刻下局部视频帧内快速运动模式进行建模

: Recurrent convolution

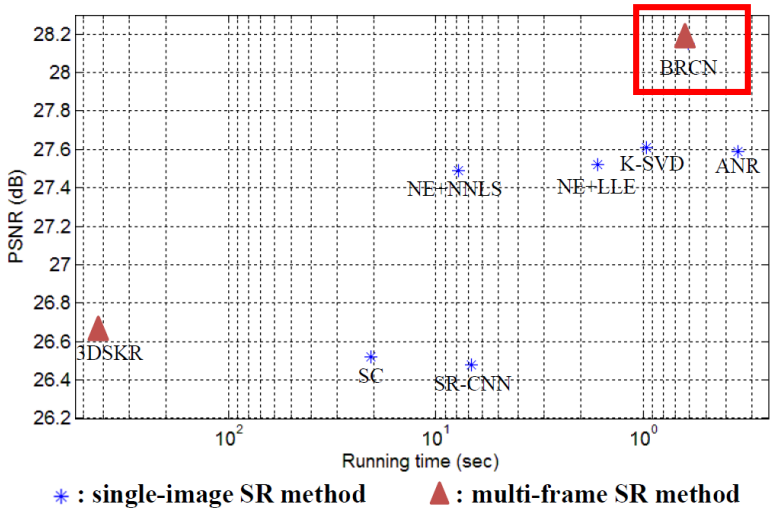
对不同视频帧之间的全局缓慢运动模式进行建模

[1] Huang et al., Video Super-resolution via Bidirectional Recurrent Convolutional Networks, **TPAMI**, 2017.

[2] Huang et al., Bidirectional Recurrent Convolutional Networks for Multi-frame Super-resolution, **NIPS**, 2015.

基于双向循环卷积网络的视频超分辨率

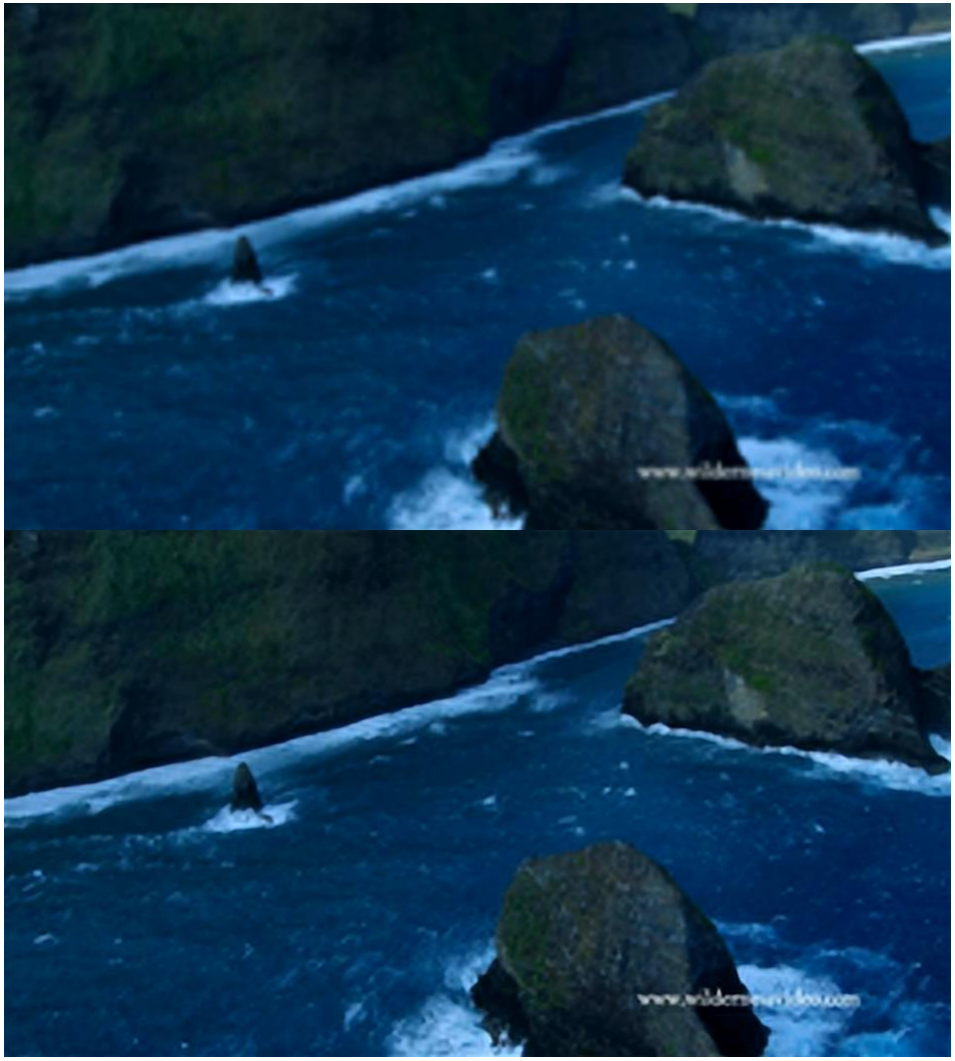
多种方法PNSR与测试时间对比



✳ : single-image SR method ▲ : multi-frame SR method

在彩色视频超分辨率上结果对比

Model	*City*	*Calendar*	*Foliage*	*Walk*	Average
			PSNR		
Bicubic	21.91	17.26	20.06	21.60	20.21
FUS [41]	21.94	17.35	19.90	21.37	20.14
Enhancer [1]	23.22	19.16	22.29	24.82	22.37
*DeepSR [30]	24.24	19.86	23.51	25.26	23.22
BRCN-Y only	23.31	19.57	22.72	25.82	22.85
BRCN-RGB	23.53	19.68	23.03	25.58	22.96
			SSIM		
Bicubic	0.4220	0.4699	0.4275	0.6846	0.5010
FUS [41]	0.4251	0.4870	0.4380	0.6910	0.5103
Enhancer [1]	0.5481	0.5730	0.5702	0.7752	0.6166
*DeepSR [30]	0.6507	0.6938	0.7248	0.7714	0.7102
BRCN-Y only	0.5447	0.5985	0.6018	0.8081	0.6383
BRCN-RGB	0.5685	0.6325	0.6213	0.8247	0.6617



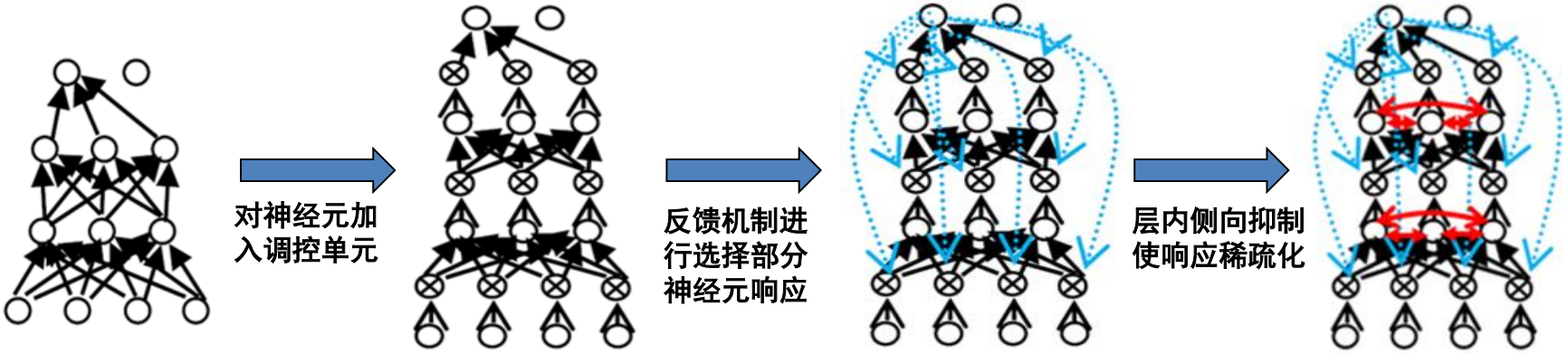
双线性插值 vs. 我们的方法

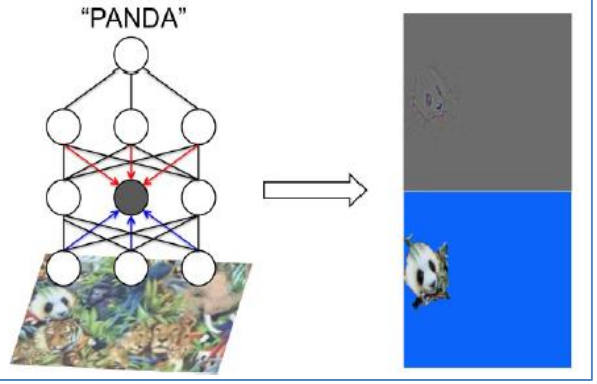
个体检测与分割

基于反馈卷积网络的弱监督目标检测和分割

任务：在弱监督条件下，查找图像视频中包含特定类别的目标

方法：提出反馈卷积网络，对人脑中反馈和侧向抑制机制建模，模拟自上而下任务驱动的视觉注意





$$x_{i,j,c} = threshold((x_{i,j,c} - mean), thd)$$

$$k_{ij,pq} = \frac{1}{\beta}\left\{\frac{1}{\beta_1}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[\frac{(d_{ij,pq}-u_1)^2}{2\sigma_1^2}\right] - \frac{1}{\beta_2}\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left[\frac{(d_{ij,pq}-u_2)^2}{2\sigma_2^2}\right]\right\}$$

$$g_{i,j,c} = sign(x_{i,j,c} - \sum k_{i,j,c,p,q,g} * x_{p,q,g})$$

侧向抑制建模

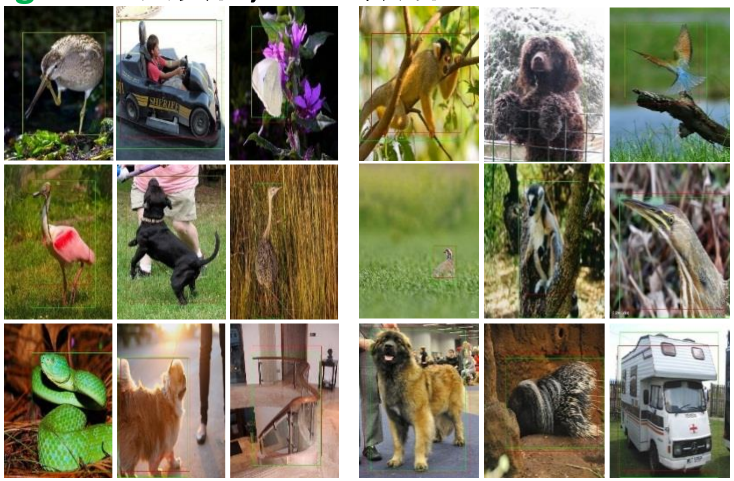
[1] Cao et al., Feedback Convolutional Neural Network for Visual Localization and Segmentation, **Submitted to TPAMI**, 2017.

[2] Cao et al., Look and Think Twice: Capturing Top-Down Visual Attention With Feedback Convolutional Neural Networks, **ICCV**, 2015.

基于反馈卷积网络的弱监督检测和目标分割

弱监督检测结果示例以及与其它方法的比较

green: 真实值, red: 预测值



	Top5 classification error	Top5 localization error
VGGnet-GAP	12.2	45.14
Backprop on VGGnet	11.4	51.46
GoogLeNet-GAP	13.2	43.00
GoogLeNet	11.3	49.34
Ours (without crop)	15.68	**42.82**
Ours (with reclassification)	12.95	**41.72**
Ours (with scale and crop)	9.22	**40.13**
Ours (with groundtruth)	0.00	**36.50**

弱监督目标分割结果示例以及与其它方法的比较

Accuracy of classification	PASCAL VOC 2012 val set	[4] (Img+Obj)	[14] (stage1)	EM-Adapt	CCNN [21]	MIL+ILP+SP-sppxly [23]	Ours
91.58	Background		71.7	67.2	68.5	77.2	**81.085**
86.90	aeroplane		30.7	29.2	25.5	37.3	**62.101**
86.84	bicycle		**30.5**	17.6	18.0	18.4	25.942
85.37	Bird		26.3	28.6	25.4	25.4	**51.511**
71.68	Boat		20.0	22.2	20.2	28.2	**32.521**
83.13	Bottle		24.2	29.6	36.3	31.9	**47.714**
78.77	Bus		39.2	47.0	46.8	41.6	**57.692**
90.98	Car		33.7	44.0	47.1	48.1	**50.950**
70.37	Cat		50.2	44.2	48.0	50.7	**65.086**
83.56	Chair		17.1	14.6	15.8	12.7	**20.629**
78.57	Cow		29.7	35.1	37.9	45.7	**55.586**
83.33	Diningtable		22.5	**24.9**	21.0	14.6	23.568
86.90	Dog		41.3	41.0	44.5	50.9	**54.523**
88.89	Horse		35.7	34.8	34.5	44.1	**54.601**
87.90	Motorbike		43.0	41.6	46.2	39.2	**57.329**
68.32	Person		36.0	32.1	**40.7**	37.9	38.511
83.33	Pottedplant		29.0	24.8	**30.4**	28.3	27.227
71.68	Sheep		34.9	37.4	36.3	44.0	**65.870**
96.47	Sofa		23.1	24.0	22.2	19.6	**31.230**
77.91	Train		33.2	38.1	38.8	37.6	**50.668**
82.62	tvmonitor		33.2	31.6	36.9	35.0	**40.245**
	average	32.2	33.6	33.8	35.3	36.6	**47.361**

提升
10.761%

在16个类别中得到最好结果

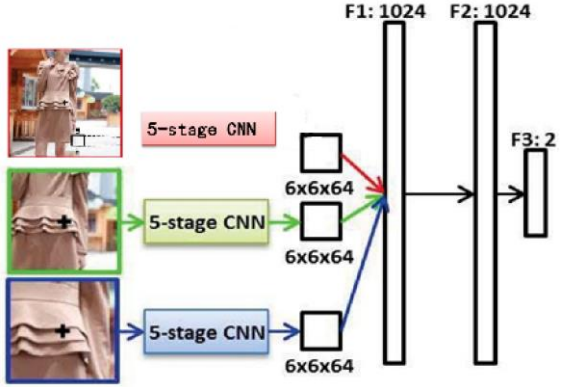
基于多尺度卷积网络的目标分割

任务： 分割视频中的前景目标

方法： 提出“点-块-图”多尺度卷积网络，融合上下文信息

各像素点单独处理

获得2013年中国云移动互联网百度人形分割大赛冠军



Team	Accuracy (%)
Second place	78.17
Third place	76.00
Forth place	75.95
Ours	**86.83**

图像块区域处理

精度略有下降，但是速度大幅提升三个数量级

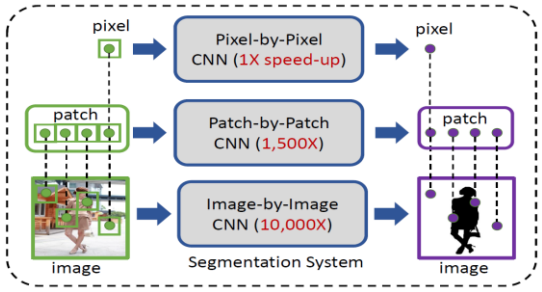
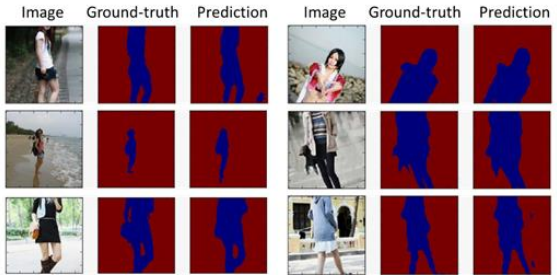


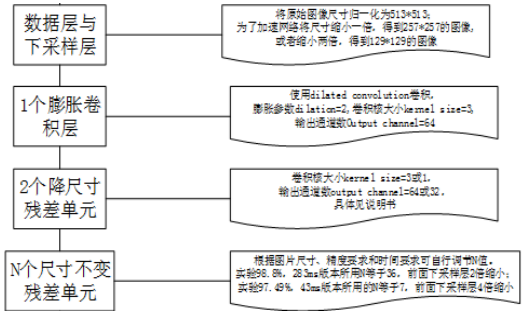
Table 1. Performance comparison of different structures

Methods	Resolution	ACC(%)	MRE	
Pixel-by-Pixel [13]	100*100	**86.83**	- -	
Simple-seg-net	48*48	62.70	147.1	
Alex-seg-net	48*48	82.12	96.2	
	112*112	80.20	490.8	
VGG-seg-net	48*48	**83.57**	**94.8**	1ms



全图整体处理

精度和速度大幅提升，在CPU上达到毫秒级



与三星、华为等公司开展合作

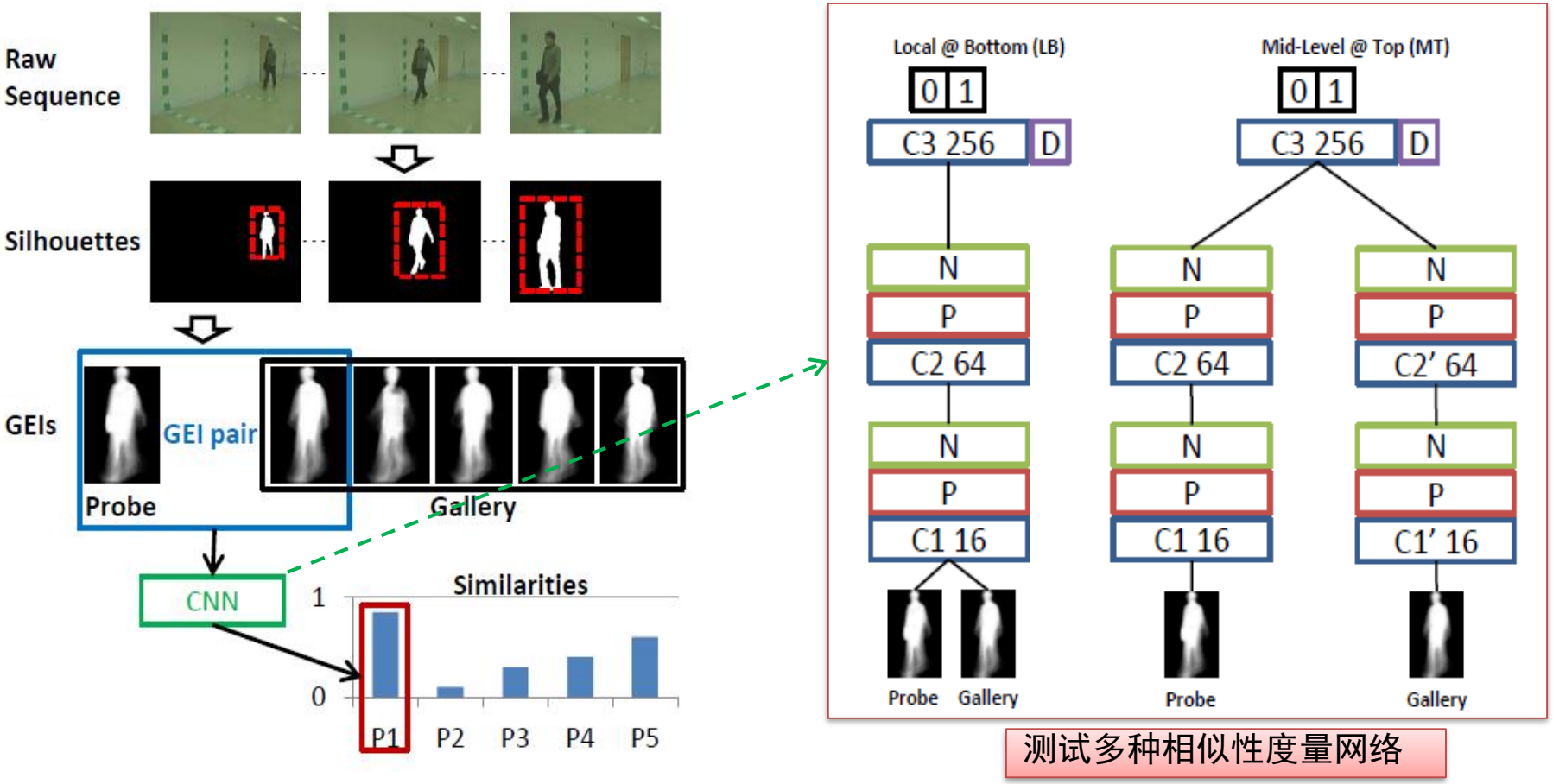
	ACC	Speed /CPU
2s	97.5%	~7000
el	97.5%	45
500model	98.8%	283

个体认证与检索

基于双通道卷积网络的步态识别

任务：对人的行走方式建模，用于远距离个体身份识别

方法：利用双通道卷积神经网络来建模不同视角下步态序列的表观变化，并度量两个序列之间的相似性



测试多种相似性度量网络

[1] Wu et al., A Comprehensive Study on Cross-view Gait based Human Identification with Deep CNNs, **TPAMI**, 2017.

基于双通道卷积网络的步态识别

最难的跨视角步态数据库CASIA-B

识别精度为90%，而之前国际最高水平仅为83%

Gallery NM #1-4	0°-180°				36°-144°		
Probe NM #5-6	0°	54°	90°	126°	54°	90°	126°
SVR [34]	–	28	29	34	35	44	45
TSVD [33]	–	39	33	42	49	50	54
CMCC [12]	46.3	52.4	48.3	56.9	-	-	-
ViDP [27]	–	59.1	50.2	57.5	83.5	76.7	80.7
Ours	**54.8**	**77.8**	**64.9**	**76.1**	**90.8**	**85.8**	**90.4**

此前国际最大的步态数据库OULP

识别精度为98%，而之前国际最高水平仅为87%

Probe angle	Gallery angle	Identical angle	
	Mean	Ours	NN [28]
55°	91.6 ± 0.2	98.8 ± 0.1	84.7
65°	92.3 ± 0.1	98.9 ± 0.2	86.6
75°	92.4 ± 0.1	98.9 ± 0.0	86.9
85°	94.8 ± 0.3	98.9 ± 0.1	85.7

大规模远距离步态识别

建成全球最大的户外步态数据库(CASIA-HT)

- 户外真实场景
- 1014人
- 3种不同着装
- 3个不同场景
- 2种行走状态
- 13个不同水平视角
- 2个不同垂直视角
- 1920x1080分辨率
- 部分红外数据

产业化

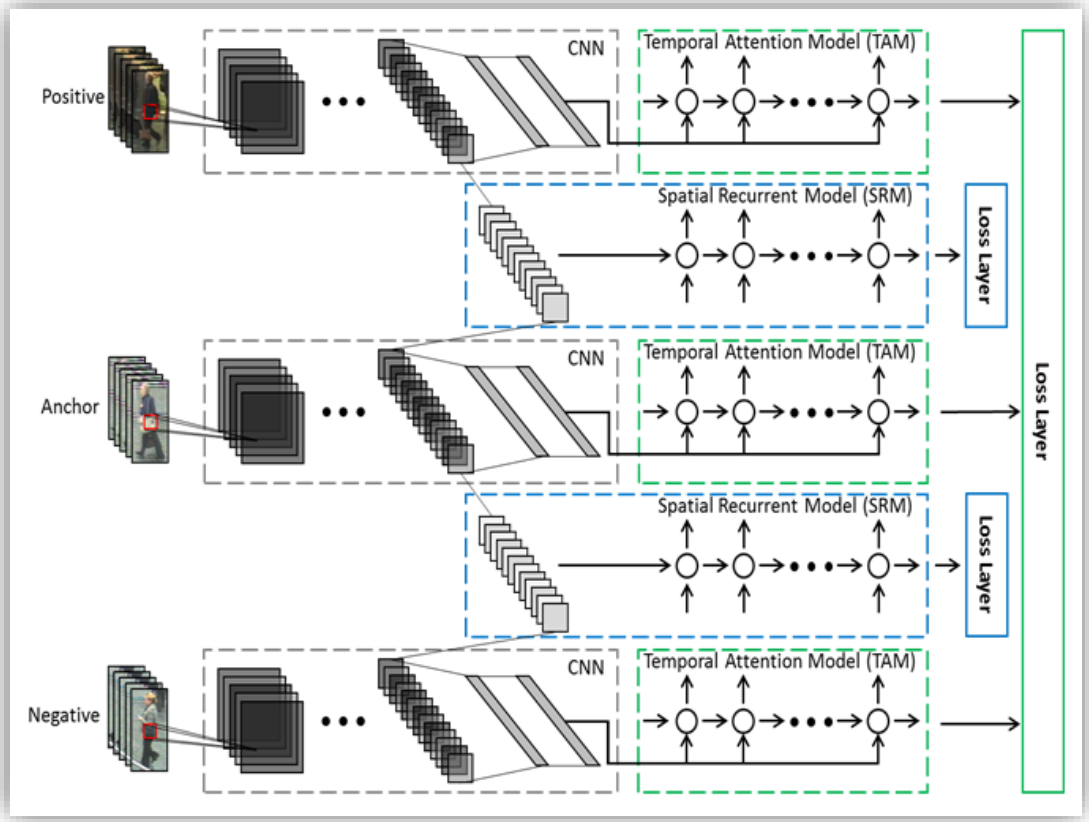
银河水滴科技(北京)有限公司

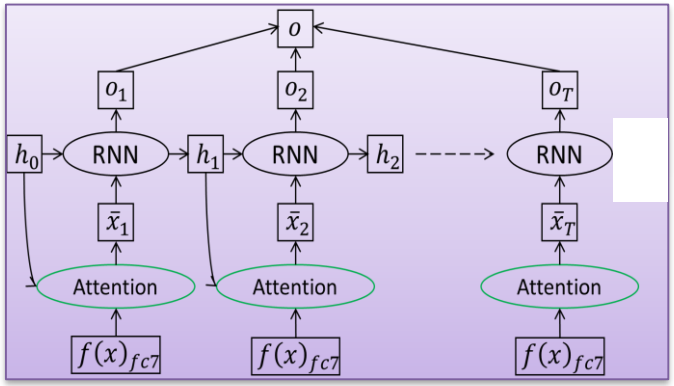
http://www.watrix.cc

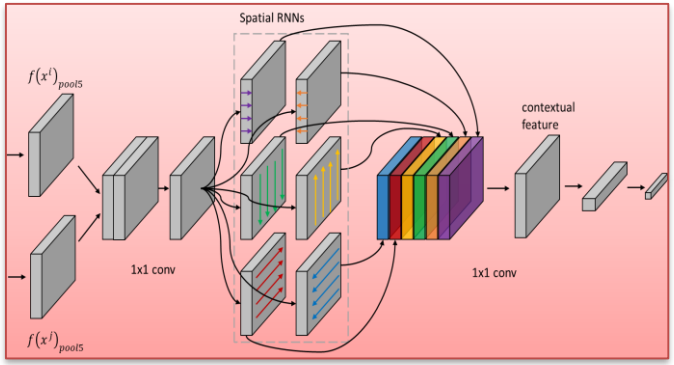
基于时空循环网络的视频行人再识别

任务：对视频中跨摄像机下行人进行再识别（或以图搜人）

方法：提出时间注意机制模型来判别式学习视频行人关键帧特征，利用循环差分模型来进行度量学习



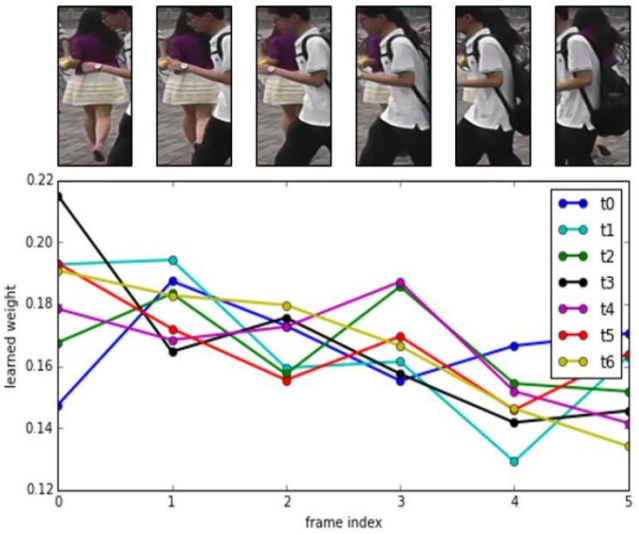




[1] Zhou et al., See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-based Person Re-identification, **CVPR**, 2017.

基于时空循环网络的视频行人再识别

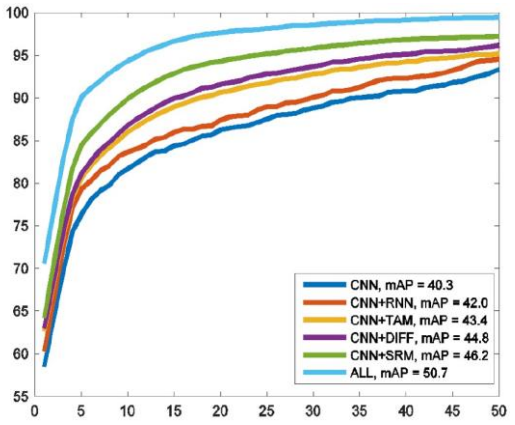
选择性关注有判别性的视频帧



检索得到最相似结果的实例

在当前最大视频行人再识别数据库上取得最好结果

Datasets	iLIDS-VID			PRID2011			MARS			
Rank@R	R = 1	R = 5	R = 20	R = 1	R = 5	R = 20	R = 1	R = 5	R = 20	mAP
Wang *et al.* [27]	34.5	56.7	77.5	37.6	63.9	89.4	-	-	-	-
Liu *et al.* [20]	44.3	71.7	91.7	64.1	87.3	92.0	-	-	-	-
Karanam *et al.* [15]	25.9	48.2	68.9	40.6	69.7	85.6	-	-	-	-
Wang *et al.* [28]	41.3	63.5	83.1	48.3	74.9	94.4	-	-	-	-
You *et al.* [36]	56.3	**87.6**	**98.3**	56.7	80.0	93.6	-	-	-	-
Mclaughlin *et al.* [21]	**58**	84	96	70	90	97	-	-	-	-
Wu *et al.* [30]	46.1	76.8	95.6	69.0	88.4	96.4	-	-	-	-
Zheng *et al.* [38]	53.0	81.4	95.1	77.3	93.5	**99.3**	68.3	82.6	89.4	49.3
Ours	55.2	86.5	97.0	**79.4**	**94.4**	**99.3**	**70.6**	**90.0**	**97.6**	**50.7**

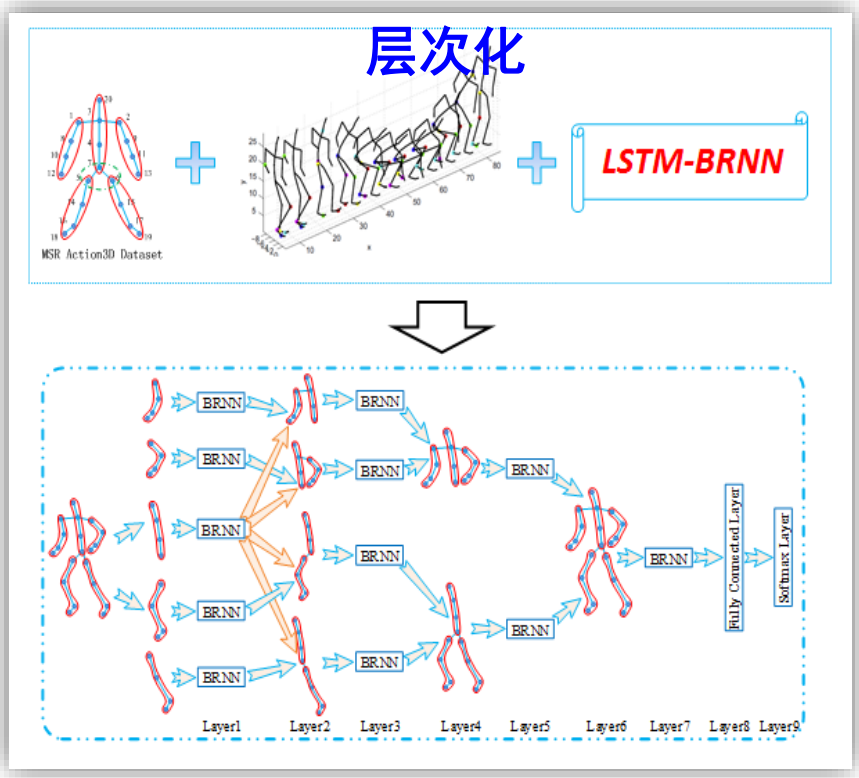


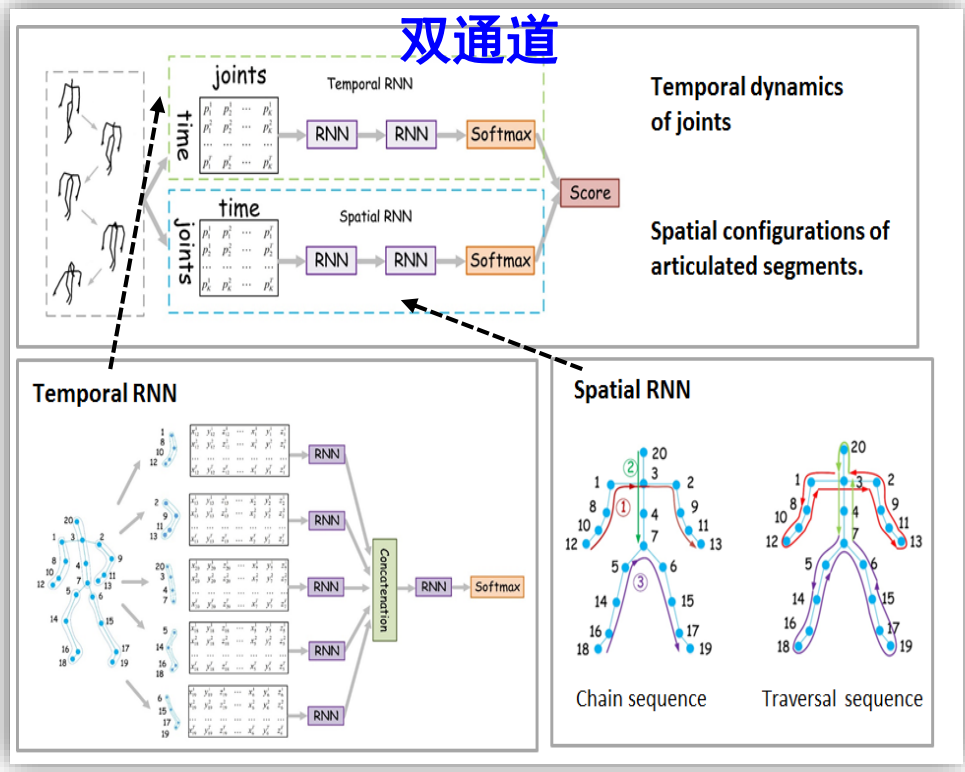
个体行为识别

基于双通道循环网络的骨架行为识别

任务：对视频中个体行为模式进行识别

方法：提出层次化/双通道递归神经网络对行为的时间动态特性和空间静态分布进行建模





[1] Wang and L. Wang, Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks, **CVPR**, 2017.

[2] Du et al., Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition, **TIP**, 2016.

[3] Du et al., Hierarchical recurrent neural network for skeleton based action recognition, **CVPR**, 2015.

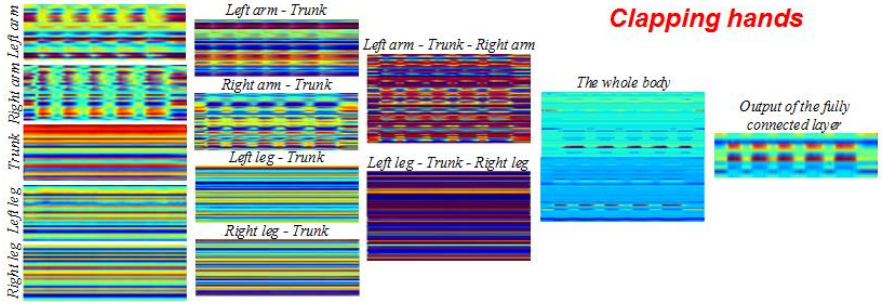
基于双通道循环网络的骨架行为识别

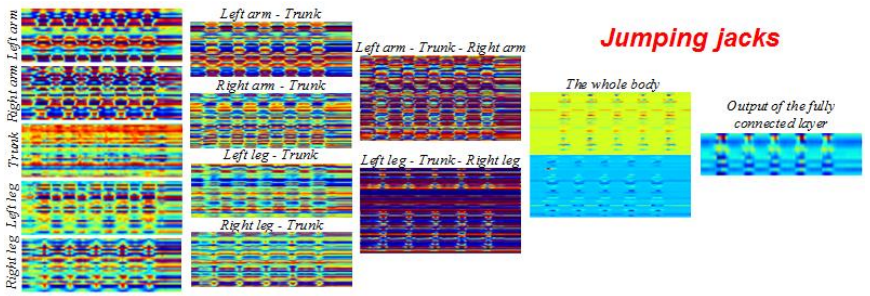
在主流行为识别数据集(NTU RGB+D&SBU)上取得了当前最好结果

Method	Cross subject	Cross view
Lie Group [7]	50.1	52.8
Skeletal Quads [2]	38.6	41.4
FTP Dynamic [3]	60.2	65.2
HBRNN [1]	59.1	64.0
Part-aware LSTM [6]	62.9	70.3
Trust Gate ST-LSTM [5]	69.2	77.7
Two-stream RNN	**71.3**	**79.5**

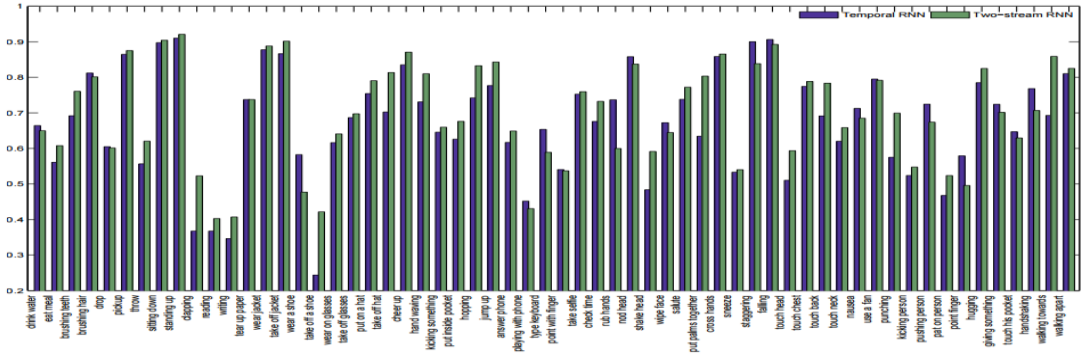
Method	Accuracy
Joint Feature [8]	80.3
Joint Feature [4]	86.9
HBRNN [1]	80.4
Deep LSTM [9]	86.0
Co-occurrence LSTM [9]	90.4
Trust Gate ST-LSTM [5]	93.3
Two-stream RNN	**94.8**

不同行为模式在网络隐含层响应的可视化

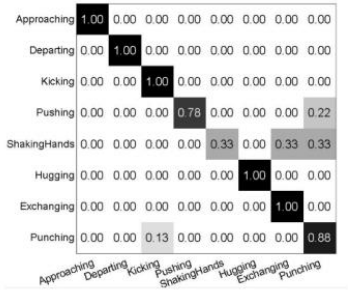




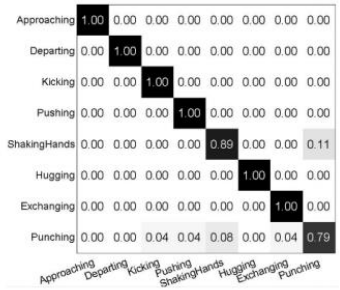
不同类别的精度提升情况



混淆矩阵对比



(a) Temporal RNN



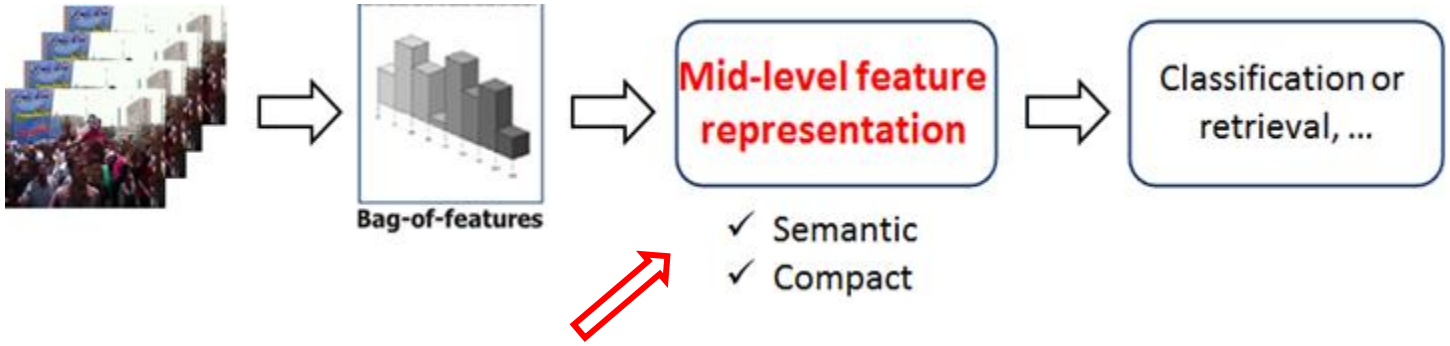
(a) Two-stream RNN

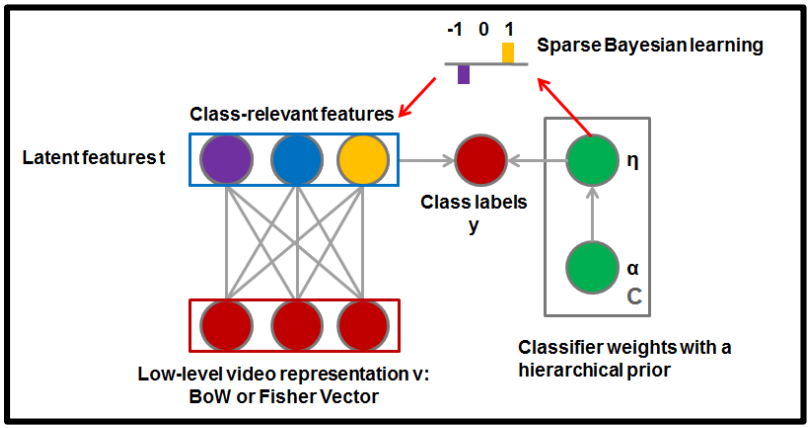
群体事件识别

基于类相关受限玻尔兹曼机的群体事件识别

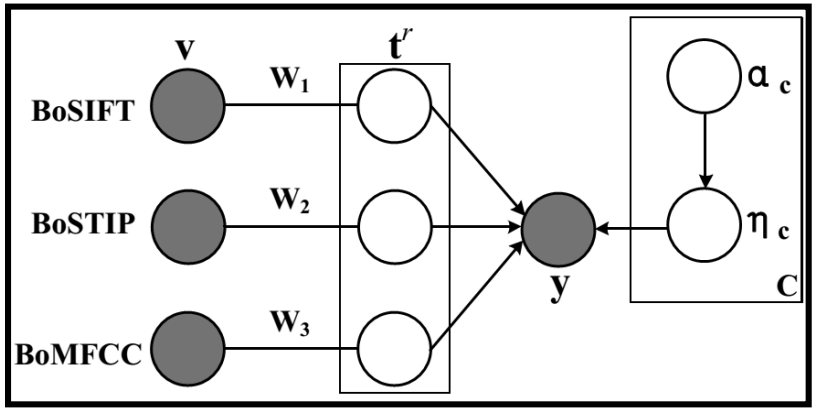
任务：处理视频中复杂群体行为与事件识别

方法：提出类相关受限玻尔兹曼机来联合学习有判别力的中层特征表示和稀疏分类器，解决语义鸿沟和稀少标记数据等问题





多模态扩展



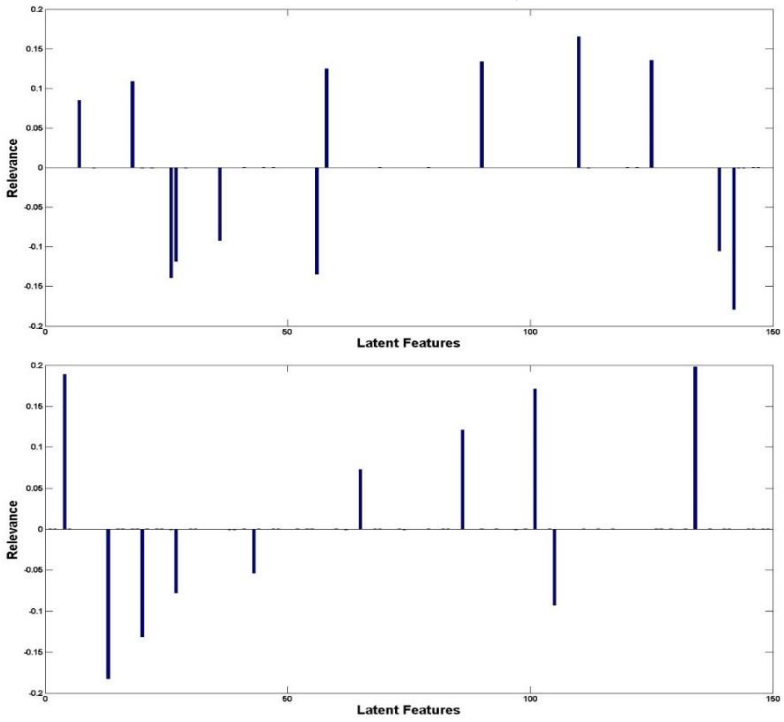
[1] Zhao et al., Learning Relevance Restricted Boltzmann Machine for Unstructured Group Activity and Event Understanding, **IJCV**, 2016.

[2] Zhao et al., Relevance Topic Model for Unstructured Social Group Activity Recognition, **NIPS**, 2013.

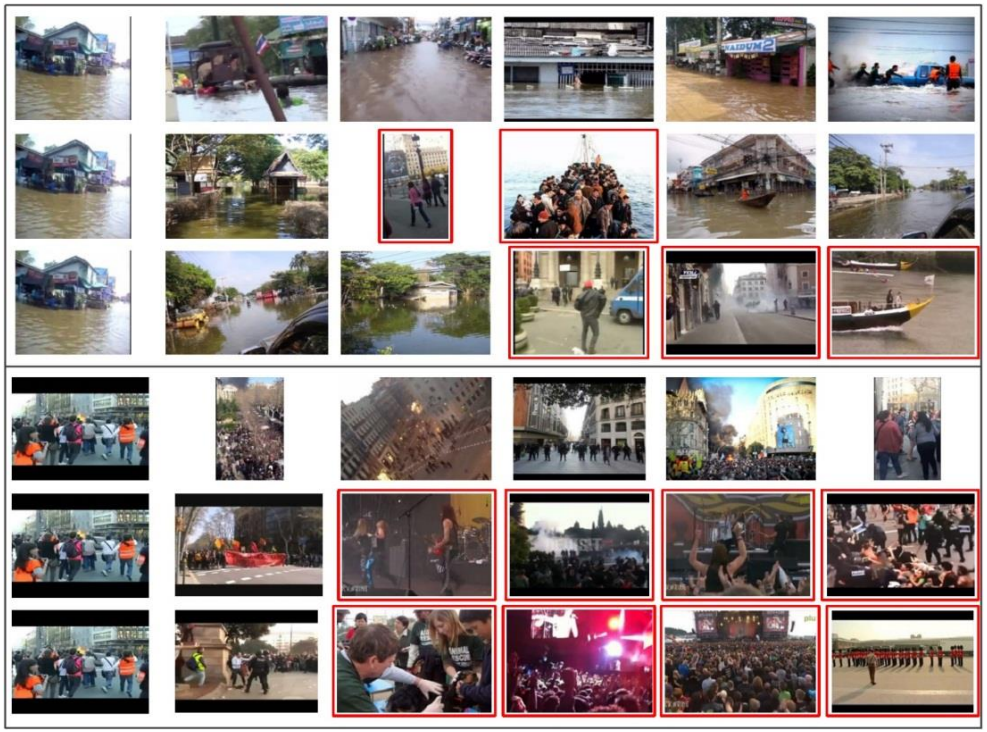
基于类相关受限玻尔兹曼机的群体事件识别

模型所关注的、与事件相关的感兴趣位置

学习到类相关的稀疏特征表示



实际检索结果的对比

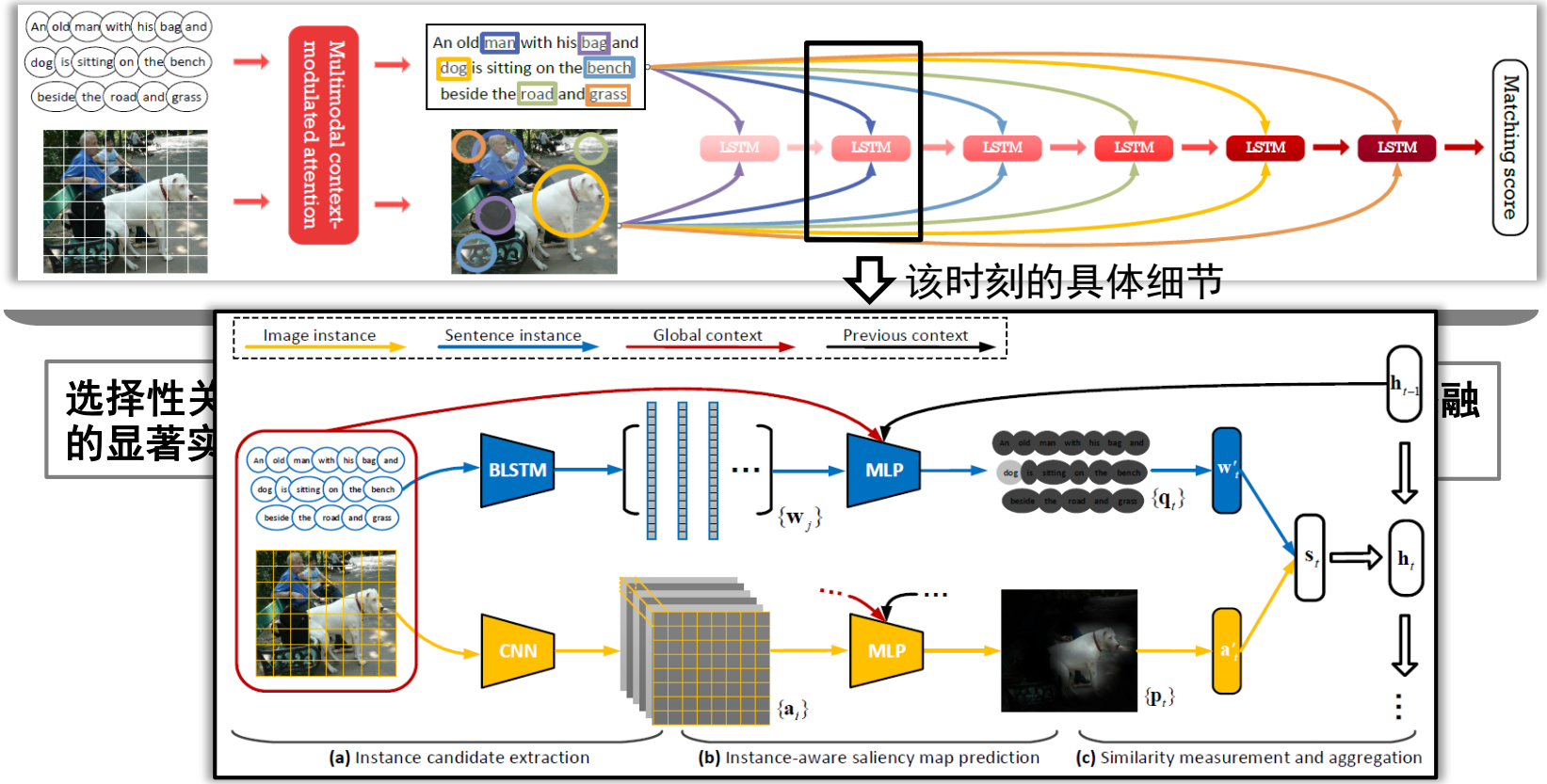


高层语义交互

基于选择式多模态循环网络的视觉语义匹配

任务：视频中基于语义属性搜索目标、视频描述生成等任务

方法：以图形文本为例研究，提出选择式匹配循环网络，从冗余信息中选择性关注和匹配显著信息



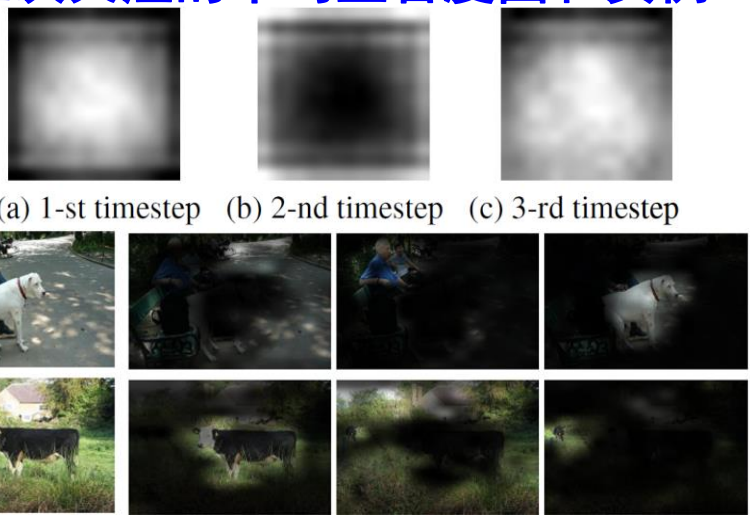
[1] Huang et al., Instance-aware Image and Sentence Matching with Selective Multimodal LSTM, **CVPR**, 2017.

基于选择式多模态循环网络的视觉语义匹配

在MSCOCO数据集上的跨模态检索结果

Method	Image Annotation				Image Retrieval				Sum
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
STD†* [16]	33.8	67.7	82.1	3	25.9	60.0	74.6	4	344.1
m-RNN [22]	41.0	73.0	83.5	2	29.0	42.2	77.0	3	345.7
FV†* [17]	39.4	67.9	80.9	2	25.1	59.8	76.6	4	349.7
DVSA [14]	38.4	69.9	80.5	**1**	27.4	60.2	74.8	3	351.2
MNLM [15]	43.4	75.7	85.8	2	31.0	66.7	79.9	3	382.5
m-CNN* [21]	42.8	73.1	84.1	2	32.6	68.6	82.8	3	384.0
RNN+FV† [19]	40.8	71.9	83.2	2	29.6	64.8	80.5	3	370.8
OEM [30]	46.7	-	88.9	2	37.9	-	85.9	**2**	-
DSPE+FV† [32]	50.1	79.7	89.2	-	39.6	75.2	86.9	-	420.7
Ours:									
sm-LSTM-mean	33.1	65.3	78.3	3	25.1	57.9	72.2	4	331.9
sm-LSTM-att	36.7	69.7	80.8	2	29.1	64.8	78.4	3	359.5
sm-LSTM-ctx	39.7	70.2	84.0	2	32.7	68.1	81.3	3	376.0
sm-LSTM	52.4	81.7	90.8	**1**	38.6	73.4	84.6	**2**	421.5
sm-LSTM*	**53.2**	**83.1**	**91.5**	**1**	**40.7**	**75.8**	**87.4**	**2**	**431.8**

三次关注的平均显著度图和实例



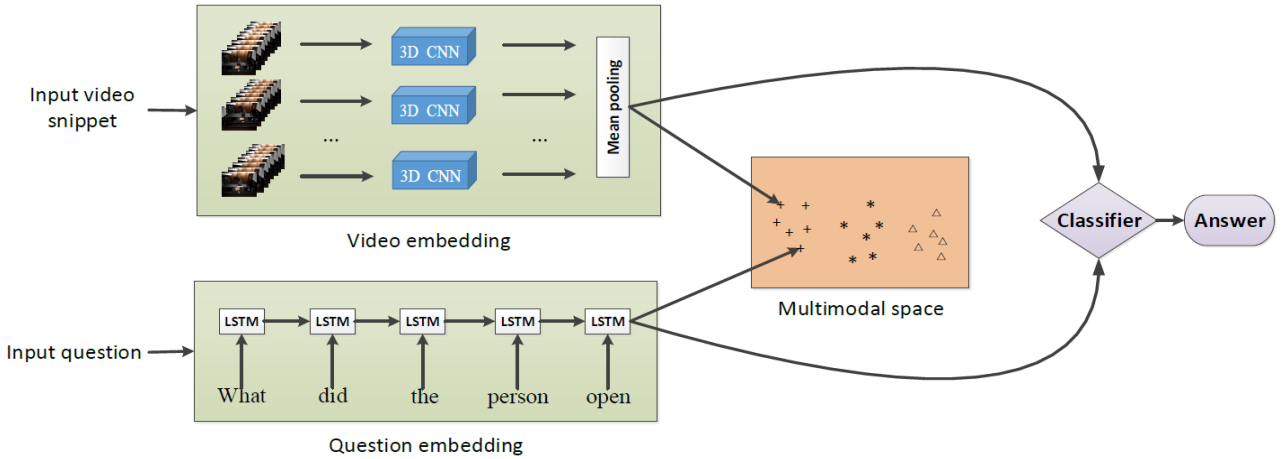
(a) 1-st timestep (b) 2-nd timestep (c) 3-rd timestep

图像文本匹配实例

基于深度语义视觉嵌入网络的视频问答

任务：对复杂场景视频进行语义交互式检索或理解目标信息

方法：基于语义属性或问句，利用深度语义视觉嵌入网络，对视频中目标的表观或者行为等信息进行问答操作



将属性或句子与视频投影到共同的空间，度量其相似性

[1] Wang et al., Long Video QA Using Deep Question-answer Embedding, **Submitted to ICCV**, 2017.

基于深度语义视觉嵌入网络的视频问答

目标表观、位置等信息识别

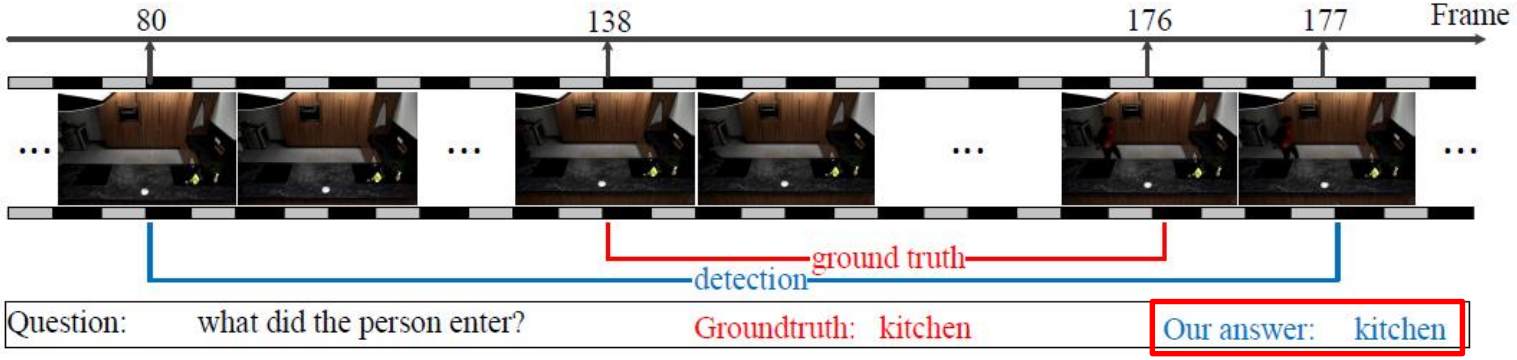
f \ q	Object	Number	Color	Location	Total
8	**14.85%**	**41.53%**	45.88%	**37.89%**	**19.59%**
16	14.50%	40.21%	**47.06%**	37.46%	19.19%
32	14.55%	30.69%	47.05%	34.79%	18.34%
64	14.50%	35.71%	45.88%	35.07%	18.62%

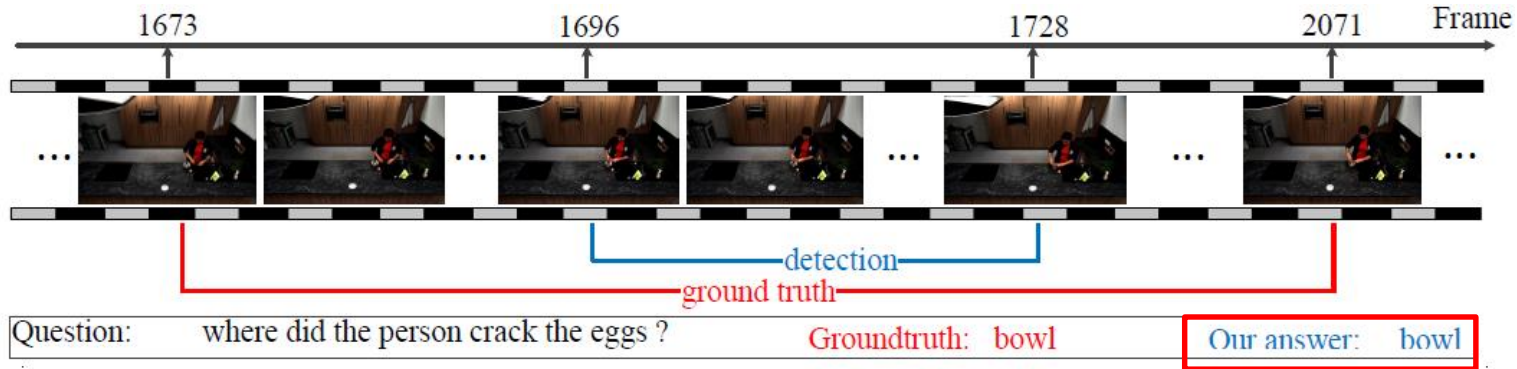
识别和检测精度都优于传统方法

f	8	16	32	64
IoU	4.54%	7.77%	4.82%	**8.29%**

f	8	16	32	64
Baseline	14.66%	14.71%	14.73%	14.56%
Ours	**19.59%**	**19.19%**	**18.34%**	**18.62%**

对目标类别、位置等信息进行问答操作实例



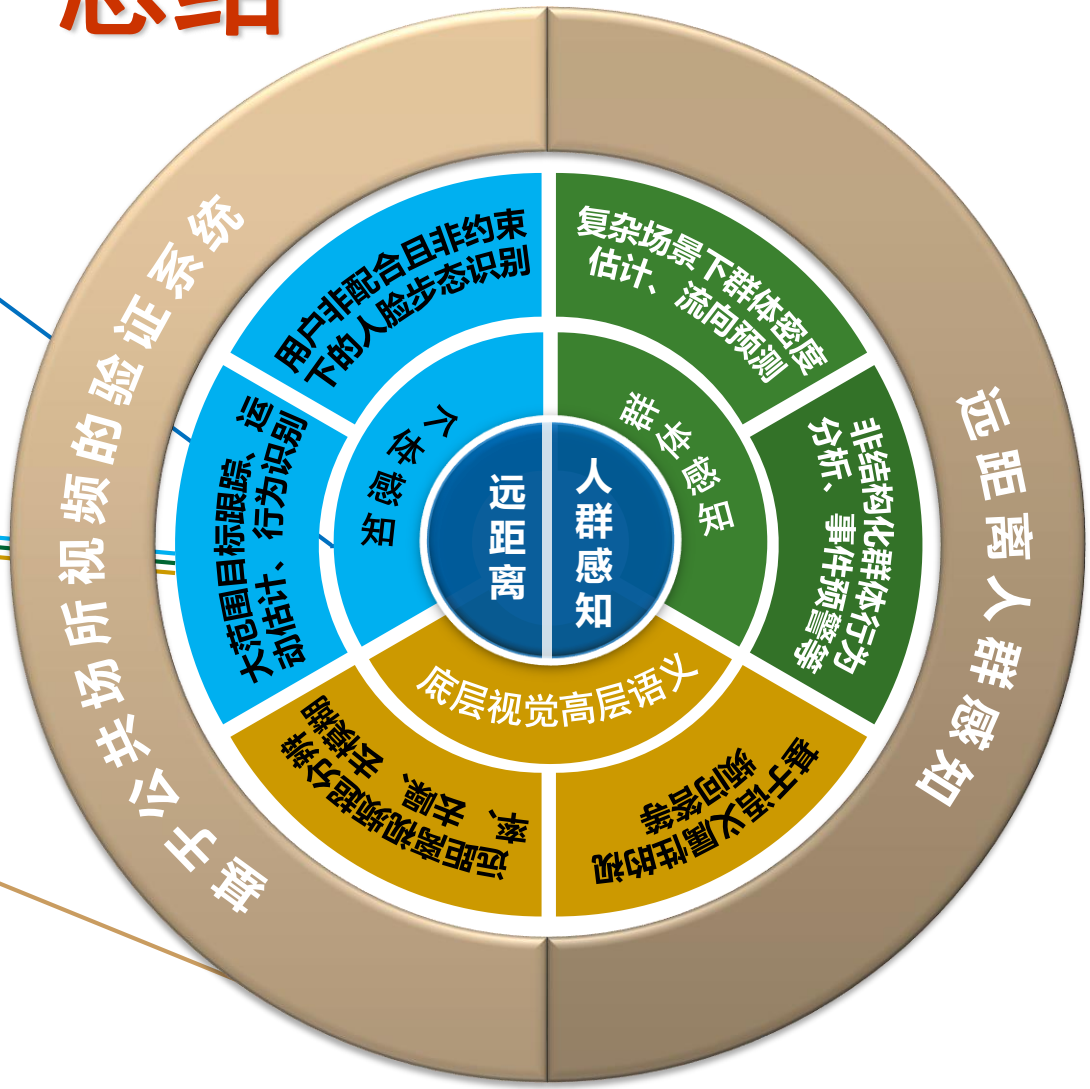


总结

面向远距离人群感知的实际需求

突破个体感知、群体感知、底层视觉分析和高层语义交互等关键技术

实现远距离人群感知视频分析系统及应用



该研究得到国家重点研发计划项目“面向大范围场景透彻感知的视觉大数据智能分析关键技术与验证系统”资助。

智能感知与计算研究中心

大数据与多模态计算研究组

联系邮箱：multimodal_comp@126.com