理解非言语交互

陈熙霖
xlchen@ict.ac.cn

中国科学院计算技术研究所

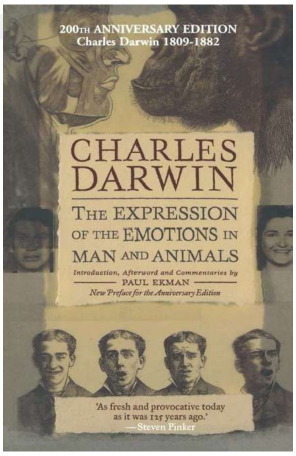
Look at the person, understand the person for serve the person better
-- A Vision for HCI and Biometrics in Next Decade

非言语交互 - Jurassic World 2

非言语交互

- 非言语交互的广泛性
  - 比言语交互更加稳定而广泛，甚至在不同类别动物之间也广泛存在
  - 人类语言数以百计

Smiling is an universal pass for human society



非言语交互传达的信息


矛盾(Conflicting)


强调(Accenting)


替代(Substituting)

对言语交互的支持(一致)程度

非言语交互通道
- 面部
  - 表情
  - 视线
  - 唇动
- 身体
  - 体势
  - 手势
  - …


重复(Repeating)

补充(Complementing)

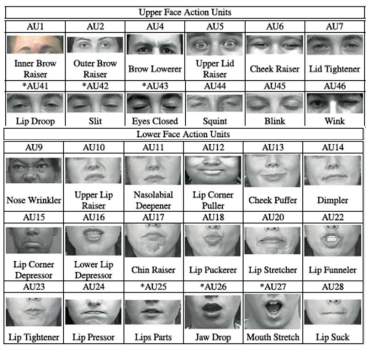
调控(Regulating)

人-机无缝交互--AI

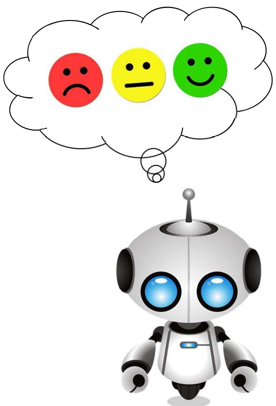
面部表情理解

- 面部表情识别

Basic emotions



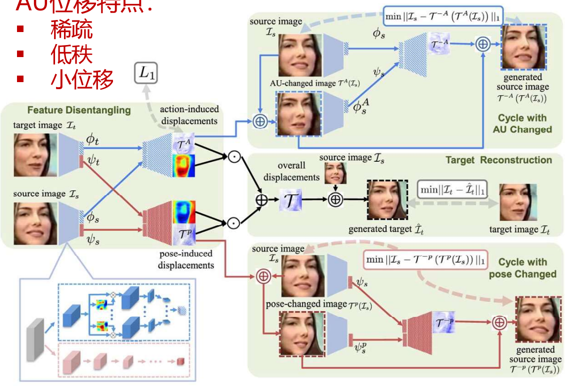
Facial Action Coding System

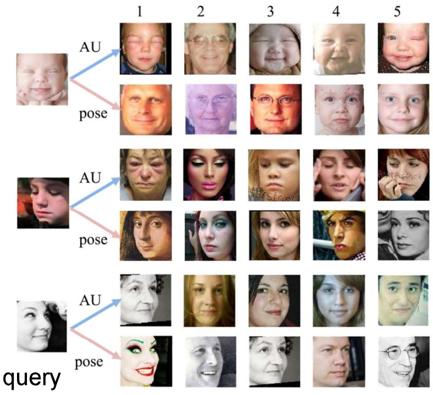


面部表情理解

- 无监督面部动作表示学习
  - Disentangle facial actions and head poses

AU位移特点：
- 稀疏
- 低秩
- 小位移



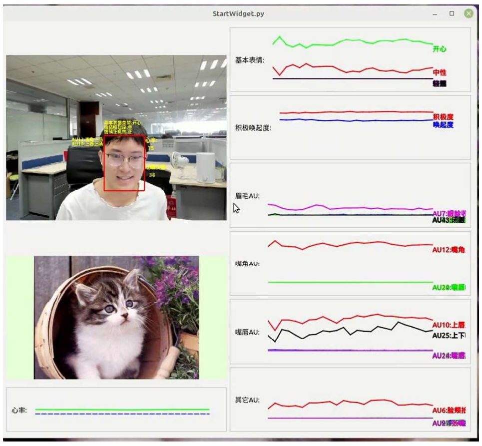


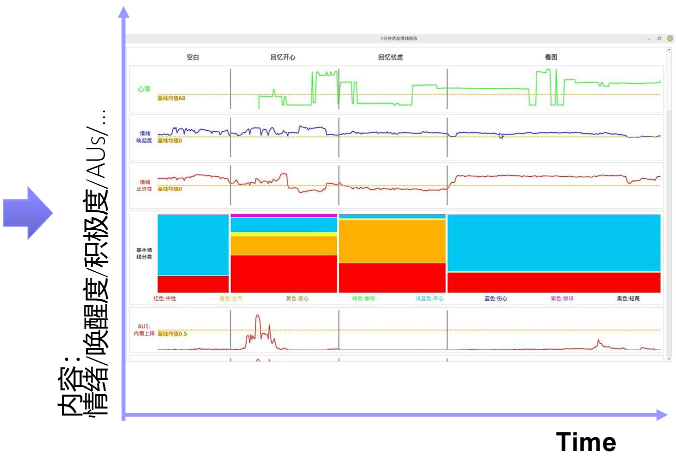
retrieved images (in descending order)

Yong Li, Jiabei Zeng, Shiguang Shan, Xilin Chen. Self-supervised Representation Learning from Videos for Facial Action Unit Detection. CVPR 2019

面部表情理解

- 通过面部表情统计来理解



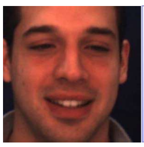


PsychoFace V1.0 developed by ICT, CAS.

从视觉-语言模型理解面部表情

The corners of the lips pulled upwards, wrinkles near the eyes, raised eyebrows, slightly widened or protruding eyes, and slight upwards or outward pushing of the lips indicate a combination of anger and disgust. Additionally, the raised upper lip or chin, slightly protruding lower lip, and flared nostrils all further intensify the emotion of anger that is evident in this face.

The slightly raised and furrowed eyebrows, along with the creased skin around the eyes, indicate a state of alertness or concentration. These actions also suggest that the individual is feeling anxious or frustrated. Additionally, the slight squint or furrow of the forehead and wrinkles on the skin around the eyes, as well as the raised eyebrows, further infer a feeling of sadness.



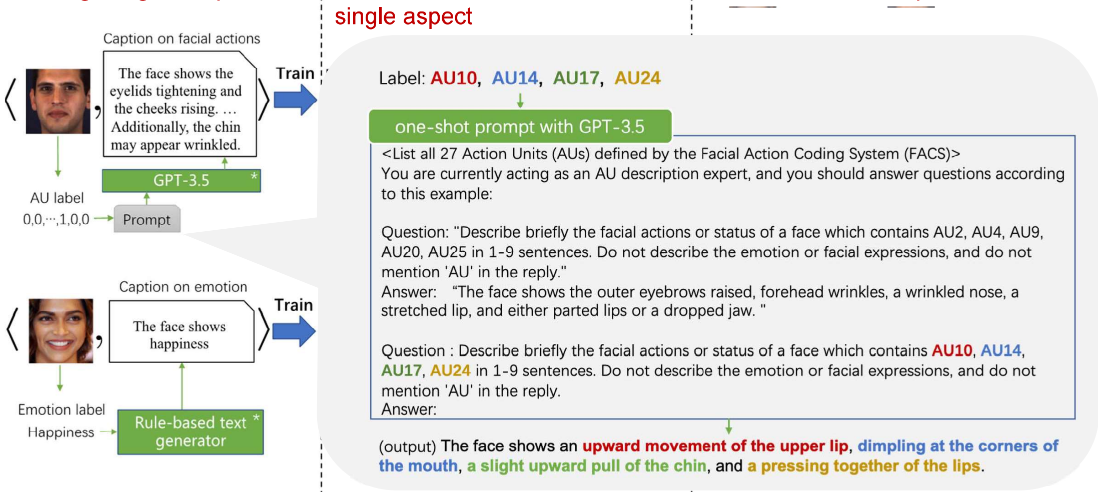
The corners of the lips pulled upwards and outward, the lower eyelids lifted, and the cheeks slightly elevated indicate happiness, while the tightened lower eyelid and horizontally stretched lips suggest a genuine sense of joy.

从视觉-语言模型理解面部表情

- Step 1: Synthesize the training image-text pairs

Step 2: Train AU-BLIP and Emot-BLIP to describe on single aspect

Step 3: Train Exp-BLIP to describe on both aspects



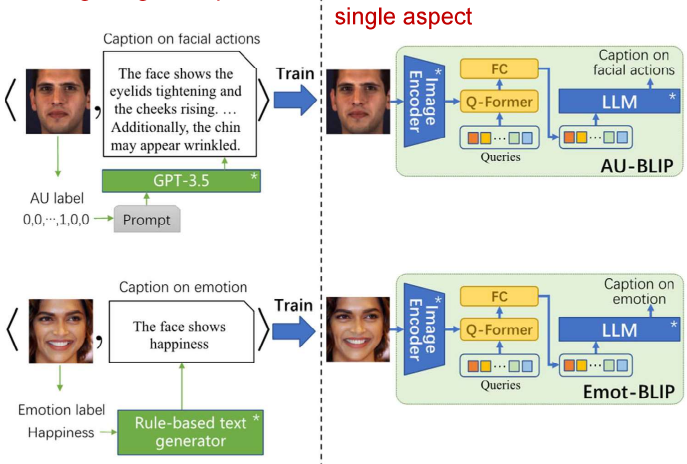
 Yujian Yuan, Jiabei, Shiguang Shan. Describe Your Facial Expressions by Linking Image Encoders and Large Language Models. BMVC2023.

从视觉-语言模型理解面部表情

- Step 1: Synthesize the training image-text pairs

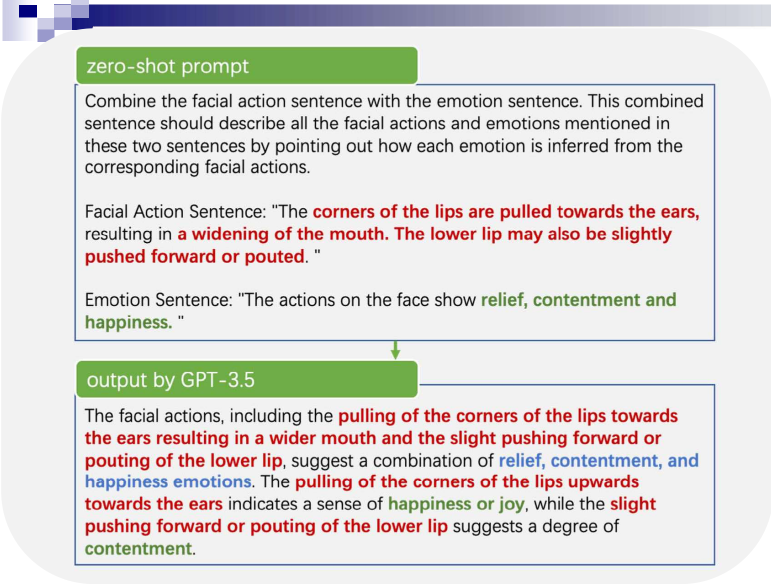
Step 2: Train AU-BLIP and Emot-BLIP to describe on single aspect

Step 3: Train Exp-BLIP to describe on both aspects

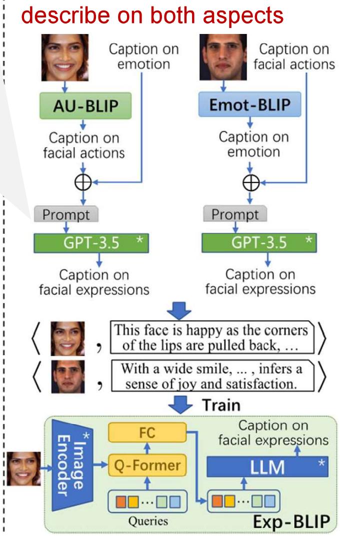


 Yujian Yuan, Jiabei, Shiguang Shan. Describe Your Facial Expressions by Linking Image Encoders and Large Language Models. BMVC2023.

从视觉-语言模型理解面部表情



Step 3: Train Exp-BLIP to describe on both aspects

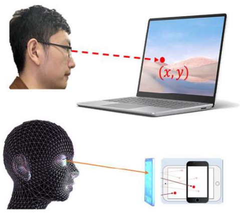


视线(Gaze)

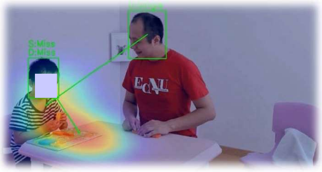
The eyes are the windows to the soul.

---*An old proverb*

3D视线指向



屏幕上的关注点

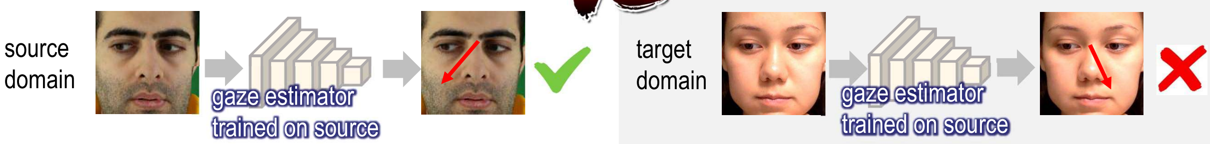


交流中的视线跟踪

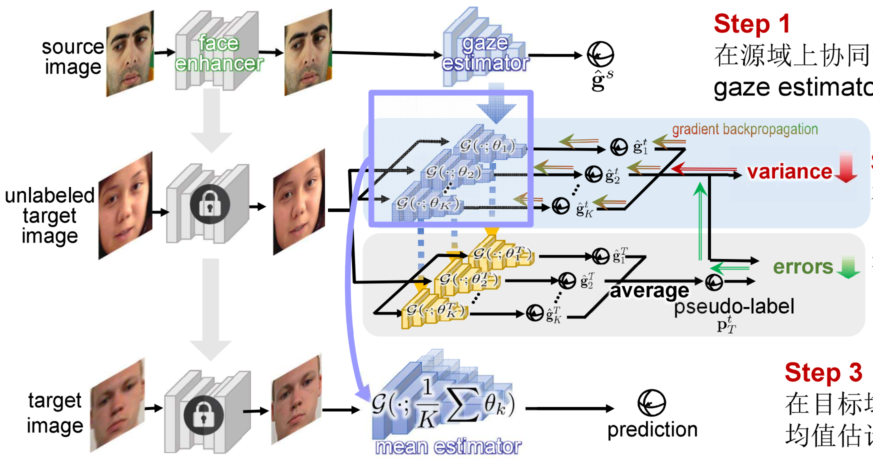
视线估计

- 挑战: 不同环境的影响

During the data collecting (source domain) **In real world application (target domain)**



视线估计



Step 1
在源域上协同训练face enhancer和gaze estimators

Step 2
在目标域数据上进行适配，目标是在伪标号监督下方差最小化

Step 3
在目标域上利用face enhancer 和均值估计器进行推断

 Xin Cai, Jiabei Zeng, Shiguang Shan, Xilin Chen. Source-free Adaptive Gaze Estimation with Uncertainty Reduction. CVPR 2023.

实验结果对比

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
Only Source	7.50	7.88	7.23	8.02
w/o source				
PureGaze [2]	7.08	7.48	9.28	9.32
PnP-GA(oma) [4]	5.65	-	6.86	-
CSA [6]	5.37	6.77	7.30	7.73
RUDA [1]	5.70	6.29	6.20	5.86
w/ source				
Gaze360 [3]	5.97	7.84	7.38	9.61
GazeAdv [5]	6.75	8.10	8.19	12.27
PnP-GA [4]	5.53	5.87	6.18	7.92
CRGA [6]	5.68	5.72	6.09	6.68
UnReGA$^-$	5.35	6.06	5.58	5.84
UnReGA	**5.11**	**5.70**	**5.42**	**5.80**

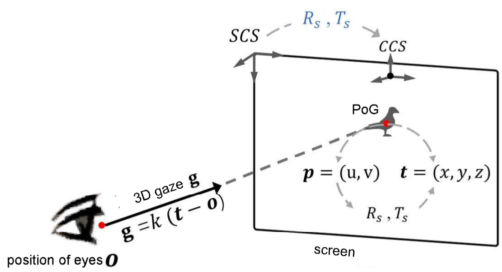
[1]. Bao et al., CVPR 2022
[2]. Cheng et al., AAAI 2022
[3]. Kellnhofer, et al., CVPR 2019
[4]. Liu et al., ICCV 2021
[5].Wang et al., CVPR 2019
[6]. Wang et al., CVPR 2022

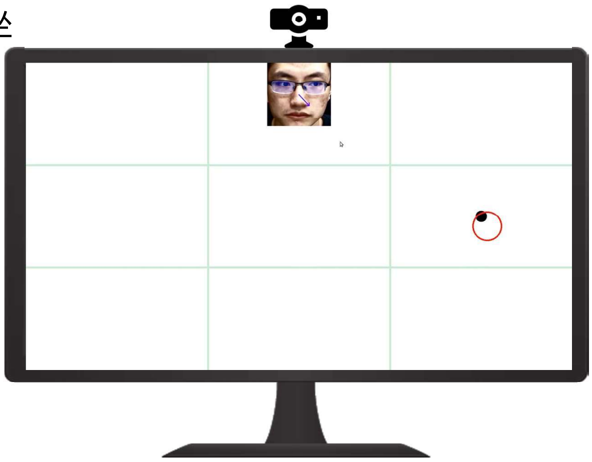
UnReGA: proposed method
UnReGA$^-$: UnReGA w/o face enhancer

D_E - ETH-XGaze, D_G - Gaze360, D_M - MPIIGaze, D_D - EyeDiap

屏幕上视线点(Point of Gaze, PoG)

- 估计PoG
  - 需要从摄像机坐标系(CCS)到屏幕坐标系(SCS)的变换
  - CCS下的空间视线方向
  - CCS下的眼睛的位置

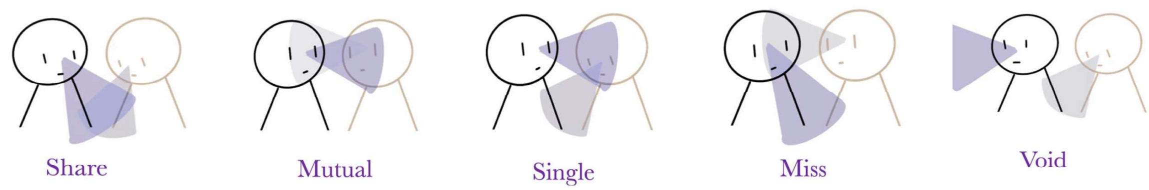




Realtime PoG estimation demo system developed by ICT, CAS.

双向视线交流模式

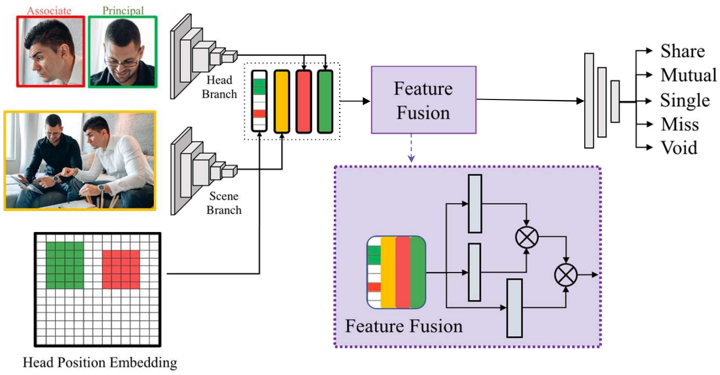
- 覆盖所有可能的个体与其伙伴的可能视线状态



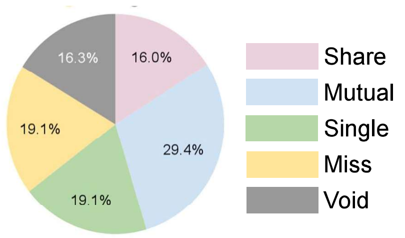
关注同一对象或空间 | 有眼睛的视线交流 | 看着伙伴，但伙伴看着别的地方 | 看着别的地方，但伙伴看着他/她 | 没有实现交流

Fei Chang, Jiabei Zeng, Qiaoyun Liu, and Shiguang Shan. Gaze Pattern Recognition in Dyadic Communication. Symposium on Eye Tracking Research and Applications (ETRA '23).

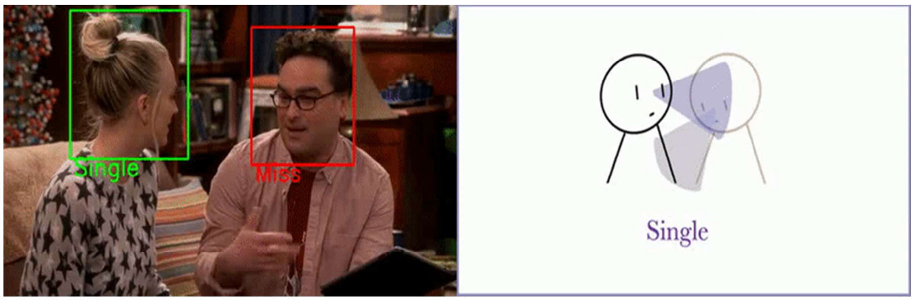
双向视线交流模式

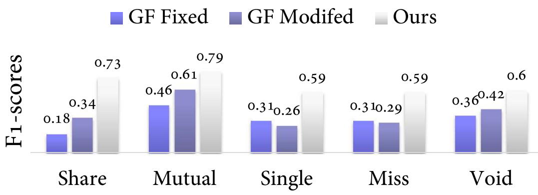


- 数据集合
  - 数据源
    - UCO-LAEO和VACATION中的部分数据
    - 370个视频片段
    - 每个片段3-15秒
  - 新的标注: 5 patterns



双向视线交流模式

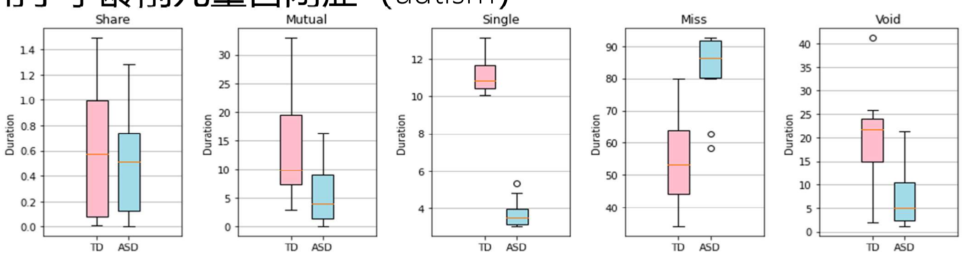




Method	Looking-At-Each-Other (AP.)			Share (Acc.)
	UCO-LAEO	AVA-LAEO	OI-MG	VideoCoAtt
LAEO-Net[Marín-Jiménez et al. 2022]	79.5	50.6	-	-
AAAI'21[Doosti et al. 2021]	65.1	72.2	70.1	-
CVPR'18[Fan et al. 2018]	-	-	-	71.4
Ours	**80.3**	**82.5**	**72.1**	**73.9**

双向视线交流模式

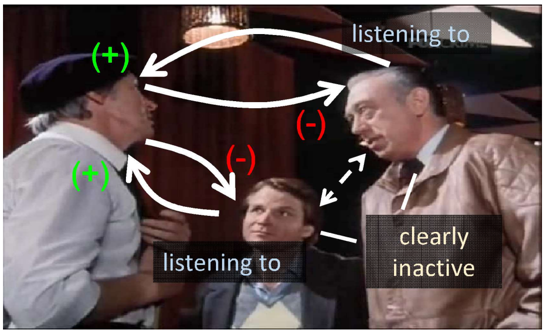
- 用于学龄前儿童自闭症（autism）

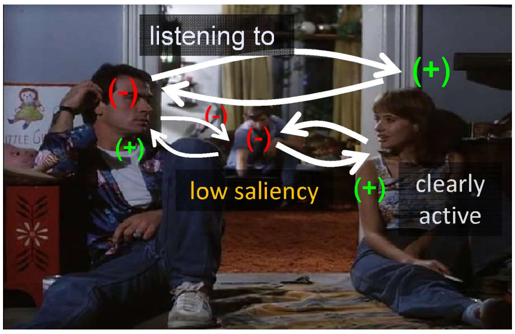


Null hypothesis	t-statistics	p-value
$H1_0$: the duration of 'Share' pattern is the same between children with and without autism	-0.46	0.66
$H2_0$: the duration of 'Mutual' pattern is the same between children with and without autism	-2.12	0.048 (*)
$H3_0$: the duration of 'Single' pattern is the same between children with and without autism	-19.00	0.0000 (***)
$H4_0$: the duration of 'Miss' pattern is the same between children with and without autism	4.54	0.0000 (***)
$H5_0$: the duration of 'Void' pattern is the same between children with and without autism	-3.07	0.006 (**)

理解视线之外的交互

- 多方交互
  - 谁是交流中的主角?
  - 谁正在谈话?
  - ……



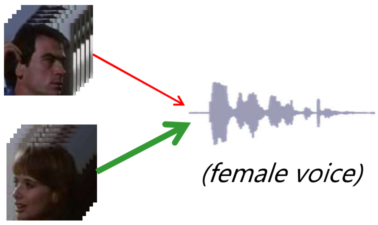


理解视线之外的交互

- 主动说话人检测
  - 从上下文线索进行判别

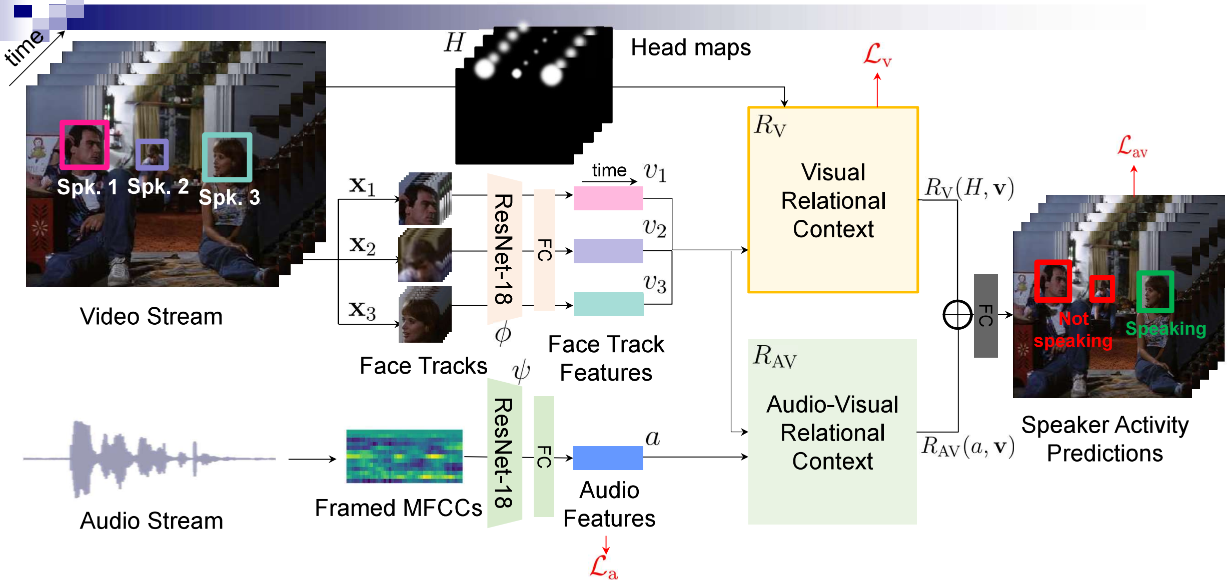

场景布局与
说话人的位置
(空间上下文)

视觉注意力
(关系上下文)



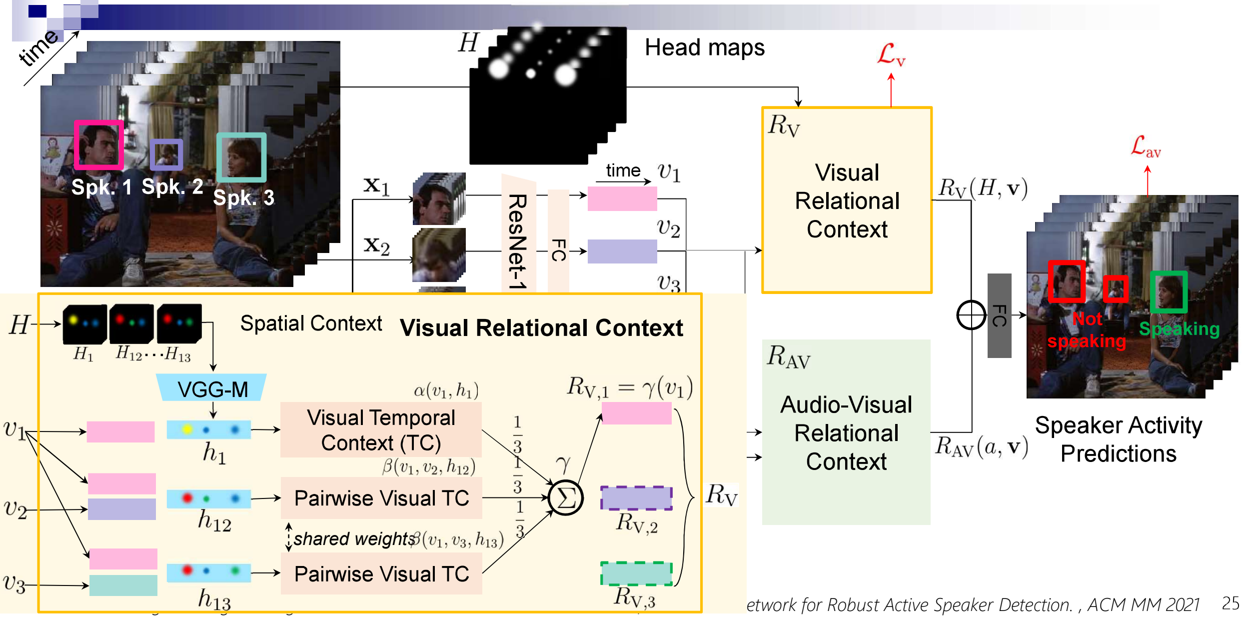
模态匹配对比
(关系上下文)

理解视线之外的交互

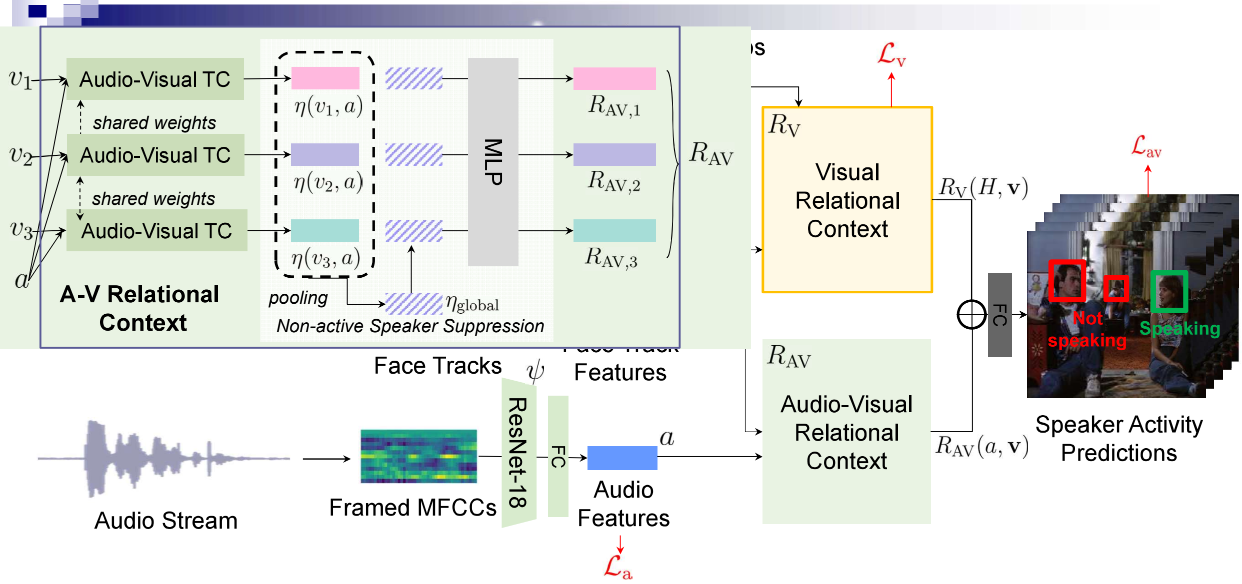


 Y. Zhang, S. Liang, S. Yang, X. Liu, Z. Wu, S. Shan, X. Chen. UniCon: Unified Context Network for Robust Active Speaker Detection. , ACM MM 2021

理解视线之外的交互



理解视线之外的交互



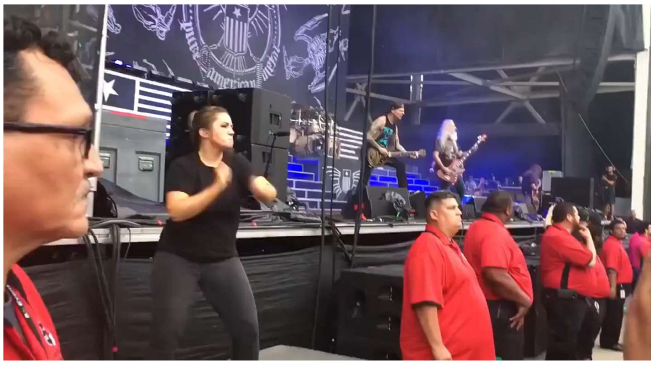
理解视线之外的交互

手势/姿势作为交互手段

- 手势/姿态传达了丰富的关于指令、情绪、态度等信息



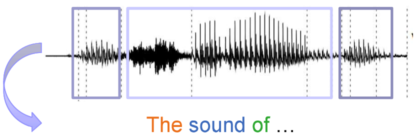


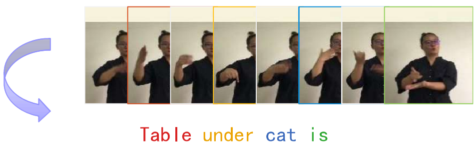



Passionate Sign Language Interpreter At Rock Gig
https://www.youtube.com/watch?v=DYoB_A8GZ08

手势/手语识别的挑战

- 自由手势的任意性与差别性
- 手语受限于细粒度标注的困难和数据的稀缺





ASR dataset scales
LibraSpeech (1000 hours)
Whisper (680,000 hours)

SLR dataset scales
Phoenix14 (12.5 hours)
CSL-Daily (23.3 hours)

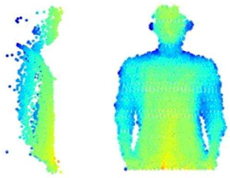
近期的一些工作

- 2020 — PointLSTM *CVPR 2020*
  探索点云在手势识别中的潜力。提出高效的 PointLSTM 以利用长期时空关系
- 2021 — VAC *ICCV 2021*
  审视 CSLR 中的迭代训练模式，揭示其中时域模块的过拟合问题。提出通过对齐视觉和上下文特征使网络更加容易训练
- 2022 — RadialCTC *ECCV 2022*
  在保留 CTC 的迭代配准机制的同时，在超球面上限制序列特征。提出一个简单的角度扰动项来控制峰值行为
- 2023 — CoSign *ICCV 2023*
  探索骨架数据在手语识别中的应用潜力。探索无标记数据在手语识别中的应用
- …

选择点云的原因

- 选择手势识别的输入模式
  - 视频
    - 计算成本高
    - 复杂的背景和照明影响
  - 点云
    - \+ 高效且稳定
    - \- 难以找到对应点

RGB Video

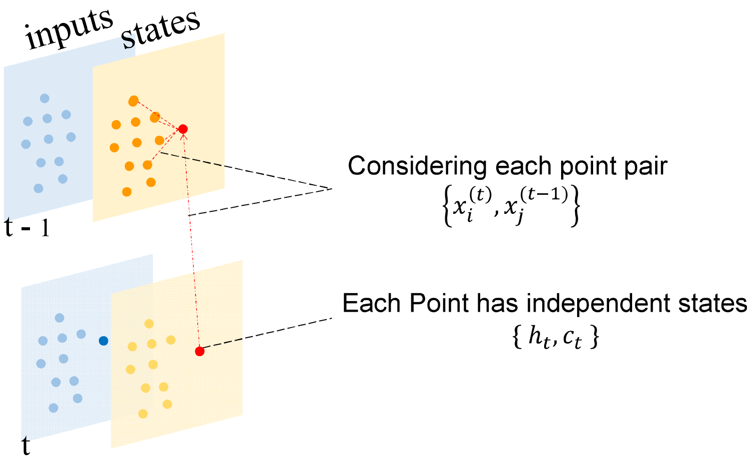


Point Cloud

Yuecong Min, Yanxiao Zhang, Xiujuan Chai, Xilin Chen, An Efficient PointLSTM for Point Clouds based Gesture Recognition, CVPR 2020

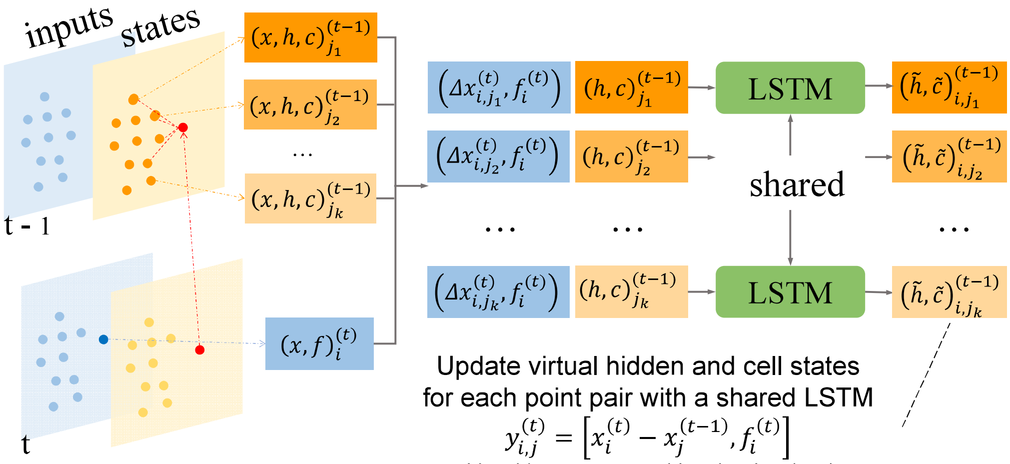
PointLSTM

- 查找过往邻点的对应点



PointLSTM

- 查找过往邻点的对应点



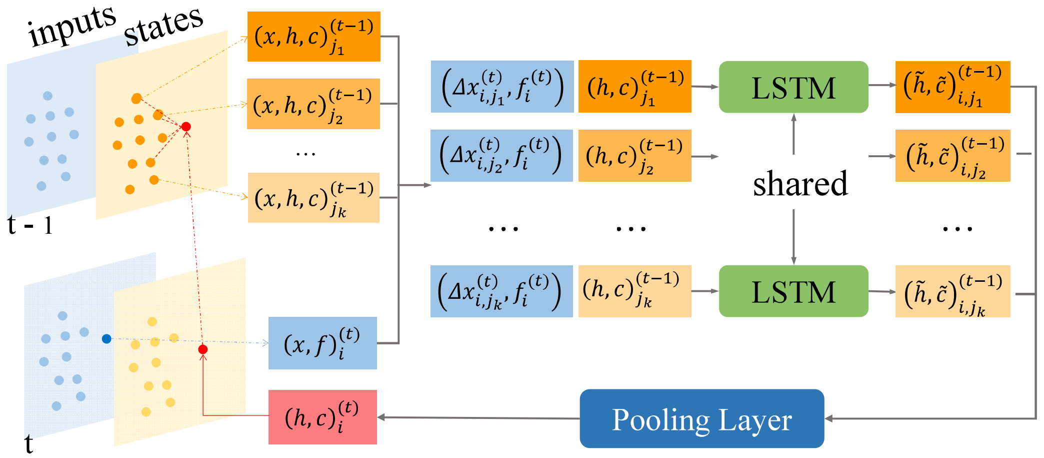
Update virtual hidden and cell states
for each point pair with a shared LSTM

$$y_{i,j}^{(t)} = \left[x_i^{(t)} - x_j^{(t-1)}, f_i^{(t)}\right]$$

$$\tilde{h}_{i,j}^{(t)}, c_{i,j}^{(t)} = \mathrm{LSTM}(y_{i,j}^{(t)}, h_j^{(t-1)}, c_j^{(t-1)})$$

PointLSTM

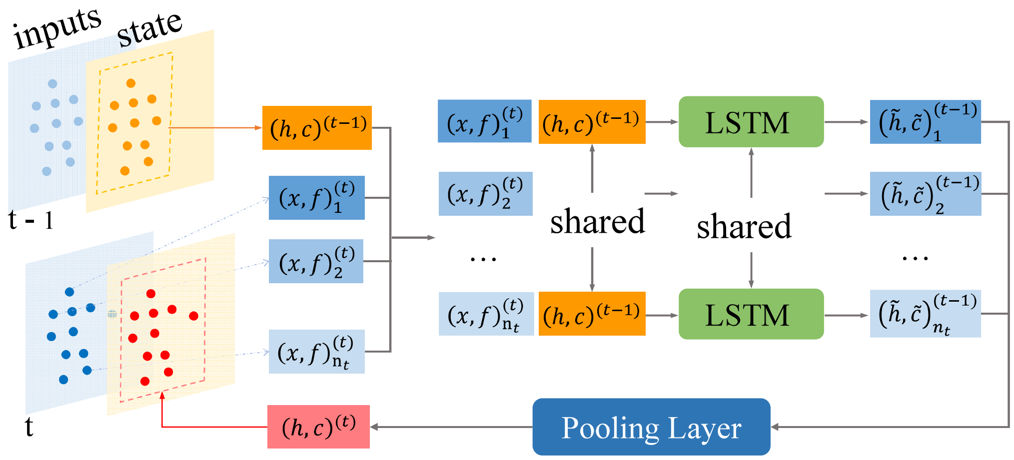
- 查找过往邻点的对应点



The final states are obtained through a pooling layer

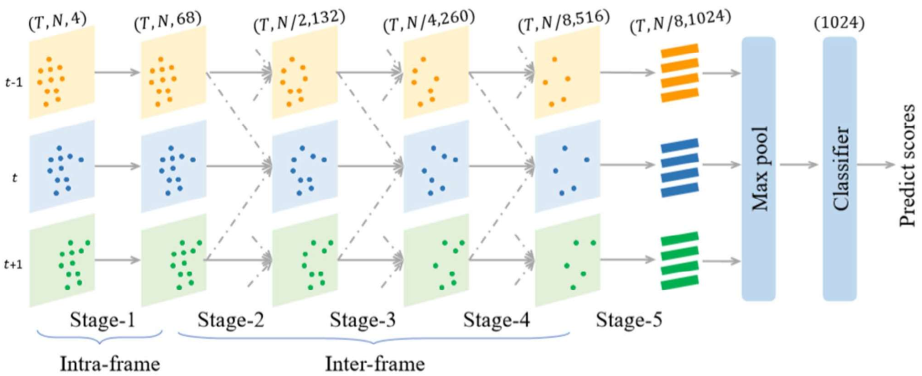
PointLSTM-PSS

- 点共享状态 (h_t, c_t)简化版



网络结构

- Baseline



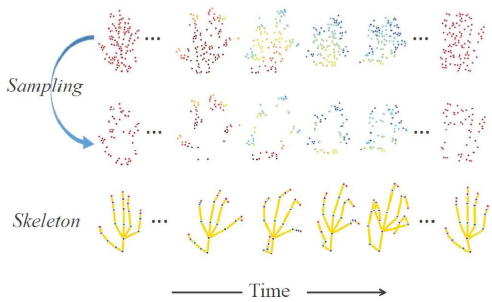
Yuecong Min, Xiujuan Chai, Lei Zhao, and Xilin Chen. Flickernet: Adaptive 3d gesture recognition from sparse point clouds. BMVC 2019.

SHREC'17上的实验

- 14 gestures × one finger or whole hand × 28 participants

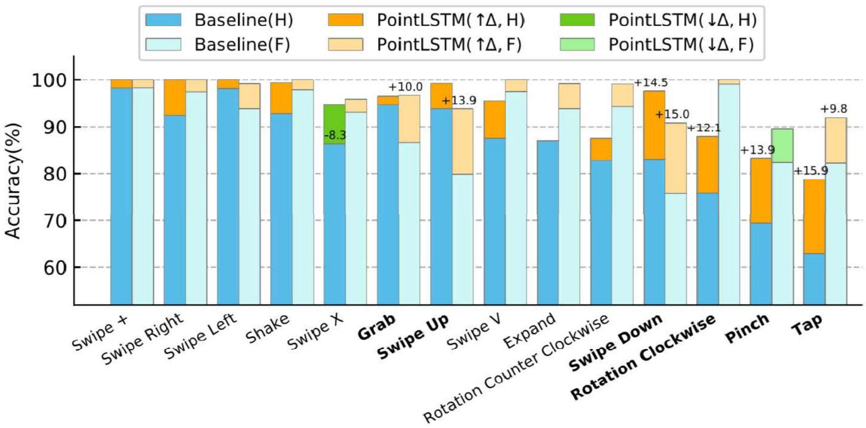
http://www-rech.telecom-lille.fr/shrec2017-hand/#gestures

Sampling 64 points from 128 points for each frame

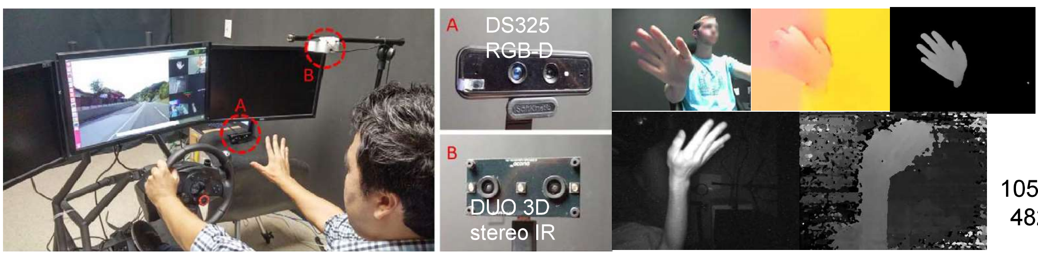


SHREC'17上的实验

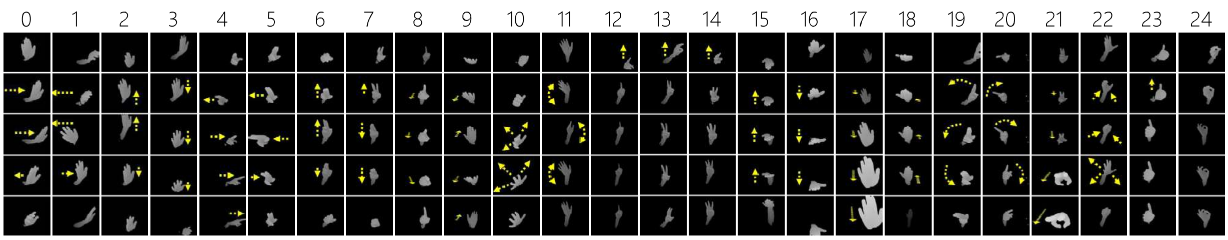
- PointLSTM-middle (94.70%) vs. Baseline (88.90%)



NvGesture上的实验



25 classes
1050 training videos
482 testing videos



P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, & J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. CVPR 2016.

NvGesture上的实验结果

Method	Modality	Accuracy
R3DCNN [1]	IR image	63.5%
R3DCNN [1]	Optical Flow	77.8%
R3DCNN [1]	Depth Video	80.3%
PreRNN [2]		84.4%
MTUT [3]		**84.9%**
R3DCNN [1]	RGB Video	74.1%
PreRNN [2]		76.5%
MTUT [3]		**81.3%**
PointNet++ [4]	Point Clouds	63.9%
FlickerNet [5]		86.3%
Baseline		85.9(±0.5)%
PointLSTM-early		**87.9(±0.7)%**
PointLSTM-PSS		87.3(±0.4)%
PointLSTM-middle		86.9(±0.6)%
PointLSTM-late		87.5(±1.0)%
Human	RGB Video	**88.4%**

[1] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, & J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. CVPR 2016.
[2] X. Yang, P. Molchanov, & J. Kautz. Making convolutional networks recurrent for visual sequence learning. CVPR 2018.
[3] M. Abavisani, H. Reza, V. Joze, & V. Patel. Improving the performance of unimodal dynamic handgesture recognition with multimodal training. ECCV 2019.
[4] C. Rui, Z. Qi, L. Yi, H. Su, & L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. NeurIPS 2017.
[5] Y. Min, X. Chai, L. Zhao, & X. Chen. Flickernet: Adaptive 3d gesture recognition from sparse point clouds. BMVC 2019.

从手势到行为动作 – MSR Action 3D

MSR Action3D*
- 20 classes × 10 subjects
- Cover various movements of arms, legs, torso and their combinations
- 567 sequences
- 20 joint locations

方法	输入模态	精度
Gram Handkel [1]	skeleton	94.74%
MeteorNet [2]	Point Clouds	88.50%
Baseline		87.62(±1.48)%
PointLSTM-early		91.78(±3.10)%
PointLSTM-PSS		90.79(±3.14)%
PointLSTM-middle		91.08(±3.43)%
PointLSTM-late		92.29(±3.09)%

[1] X. Zhang, Y. Wang, M. Gou, M. Sznaier, & O. Camps. Efficient temporal sequence comparison and classification using gram matrix embeddings on a Riemannian manifold. CVPR 2016.
[2] X. Liu, M. Yan, & J. Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. CVPR 2019.

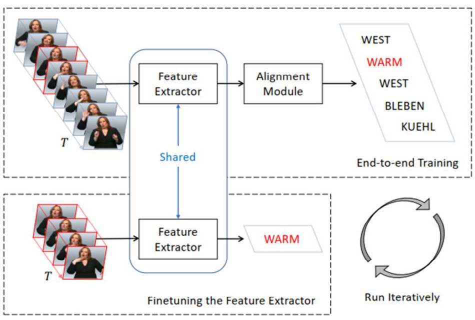
* Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. CVPRW 2010.

视频手语识别

- 捕捉精细的视觉线索
  - SLR 模型很难捕捉到正确的视觉线索
    - 迭代训练方案在 SLR 中得到广泛应用

SLR

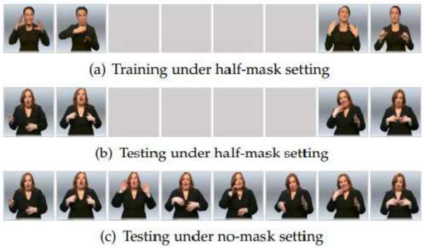
ICH OSTERN WETTER ZUFRIEDEN MITTAG TEMPERATUR SUED WARM MEIN NICHT



(a) Iterative training scheme.

视频手语识别

- 捕捉精细的视觉线索
  - SLR 模型很难捕捉到正确的视觉线索
    - 迭代训练方案在 SLR 中得到广泛应用
  - SLR 模型更倾向于根据上下文信息进行预测
    - SLR 模型即使在屏蔽一半序列的情况下也能拟合出训练集


(a) Training under half-mask setting
(b) Testing under half-mask setting
(c) Testing under no-mask setting

Method	Evaluation setting	WER on Training set (↓)	WER on Dev set (↓)
Baseline	half-mask	17.5	59.5
Baseline	no-mask	31.6	49.6
Baseline+VAC	half-mask	22.2	50.1
Baseline+VAC	no-mask	16.8	29.9

利用 VAC(Visual Alignment Constraints) 实现端到端训练

- 利用视觉特征和上下文特征之间的对应关系
- 对视觉特征引入辅助分类器
- Visual Enhancement Constraint (VEC)

$$L_{VE} = L_{CTC}^{v} = -\log p(l|x;\theta^{v})$$

Predictions from primary classifier

- Visual Alignment Constraint (VAC)

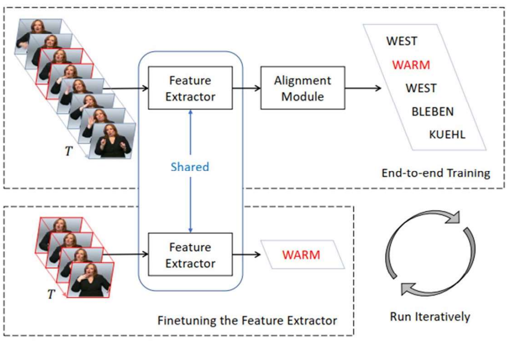
Predictions from auxiliary classifier

$$L_{VA} = \mathrm{KL}\left(\mathrm{softmax}\left(\frac{Z}{\tau}\right), \mathrm{softmax}\left(\frac{\hat{Z}}{\tau}\right)\right)$$

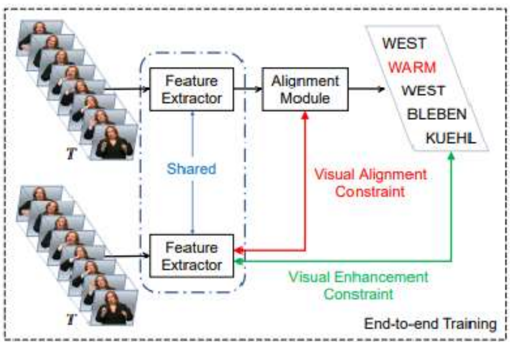
- 更简化版本的 VAC：
  共享辅助分类器和主分类器的权重

利用VAC实现端到端训练

- 迭代训练 vs. 使用 VAC 的端到端训练



(a) Iterative training scheme.



(b) End-to-end training with the proposed VCs

 Yuecong Min, Aiming Hao, Xiujuan Chai, Xilin Chen. Visual Alignment Constraint for Continuous Sign Language Recognition, ICCV 2021

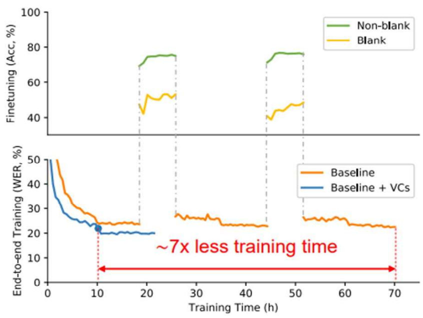
Phoenix14的实验

- 电视上的'Real-life'采集
  - 9 signers, 5.6K training samples (10.7 hours), 1081 classes

Phoenix14上的实验结果

- 利用VAC的端到端训练 vs. 迭代训练模式
  - 训练效率更高，精度更高

Iterations	Constraint	Dev	Test
1	-	23.4	24.7
2	-	22.6	22.7
3	-	22.3	23.0
-	-	23.1	24.2
-	VEC	20.7	21.0
-	VEC & VAC	**19.6**	**19.7**



Phoenix14上的实验结果

Method	Iteration	Dev	Test
SubUNet [1]		40.8	40.7
Re-Sign [2]	✓	27.1	26.8
CNN+LSTM+HMM [3]	✓	26.0	26.0
FCN [4]		23.7	23.9
DNF [5]	✓	23.1	22.9
CMA [6]	✓	21.3	21.9
STMC [7]	✓	21.1	20.7
Baseline		22.7	23.5
Baseline+VAC		**19.6**	**19.7**

[1] N. C. Camgoz, S. Hadfield, O. Koller, R. Bowden. "Subunets: End-to-end hand shape and continuous sign language recognition." ICCV, 2017.

[2] O. Koller, S. Zargaran, H. Ney. "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs." CVPR, 2017.

[3] O. Koller, N. C. Camgoz, H. Ney, R. Bowden. "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos." TPAMI, 2019.

[4] K. L. Cheng, Z. Yang, Q. Chen, Y.-W. Tai. "Fully Convolutional Networks for Continuous Sign Language Recognition." ECCV, 2020.

[5] R. Cui, H. Liu, C. Zhang. "A deep neural framework for continuous sign language recognition by iterative training." TMM, 2019.

[6] J. Pu, W. Zhou, H. Hu, H. Li. "Boosting Continuous Sign Language Recognition via Cross Modality Augmentation." ACM MM, 2020.

[7] H. Zhou, W. Zhou, Y. Zhou, H. Li. Spatial-temporal multi-cue network for continuous sign language recognition, AAAI 2020.

Phoenix14上的实验结果

Training Set

Dev. Set

Ref : UND TAG BLEIBEN KUEHL SECHS GRAD BAYERN IX REGION AUCH S+H IX AUCH ABER
Baseline : DANN TAG BLEIBEN LUFT SECHS GRAD BAYERN IX REGION AUCH SCHLESWIG H IX AUCH ABER
VAC : DANN TAG BLEIBEN SECHS GRAD BAYERN IX REGION AUCH S H AUCH ABER


Baseline


VAC

Pay more attention on facial expression with VAC

几点思考

- 非言语通道的理解对于未来自然的人-机共生环境是极其重要的
- 与言语交流相比，非言语互动的不确定性更大，更具挑战性
- 非言语交流应成为生物特征识别和人机交互领域的关注焦点

Look at the person, understand the person for serve the person better
-- A Vision for HCI and Biometrics in Next Decade

感谢我的同事和学生

- 山世光
- 曾加贝
- 杨双
- 何明捷
- 柴秀娟
- 闵越聪
- 李勇
- 蔡昕
- 常菲
- 张远航

Thanks for your attention

Look at the person, understand the person for serve the person better
-- A Vision for HCI and Biometrics in Next Decade