Lec9 Action Recognition

·AIPKU·

人工智能引论实践课 计算机视觉小班

主讲人：刘家瑛

部分讲义来源与致谢

1. Yong Du, Wei Wang, Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. CVPR 2015.
2. Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. AAAI 2016.
3. Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, Alex C Kot. Global context-aware attention LSTM networks for 3D action recognition. CVPR 2017.
4. Pichao Wang, Zhaoyang Li, Yonghong Hou, Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks. ACM MM 2016.
5. Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. CVPR 2017.
6. Karen Simonyan, Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. NIPS 2014.
7. Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell. Long-Term recurrent convolutional networks for visual recognition and description. CVPR 2015.

Video in Big Data Era

- **Videos in Internet**
  - Over a billion users on YouTube
  - A billion hours of videos each day

- **Surveillance Videos**
  - 176 million in China in 2017
  - Expected 626 million by 2020

Video in Big Data Era

- **Videos in Internet**
  - Over a billion users on YouTube
  - A billion hours of videos each day

- **Surveillance Videos**
  - 176 million in China in 2017
  - Expected 626 million by 2020

Huge number of videos contains **Human Action**

→ **Video Action Analytics**

Various Applications

Surveillance

Retrieval

HCI

Home Care

Related Topics on Action Analytics

- **Action Recognition**
  - Trimmed video → Action class
  - What?

- **Action Detection**
  - Untrimmed video → Action class & Start/end time of various actions
  - What & When?

Action Recognition

Action Detection

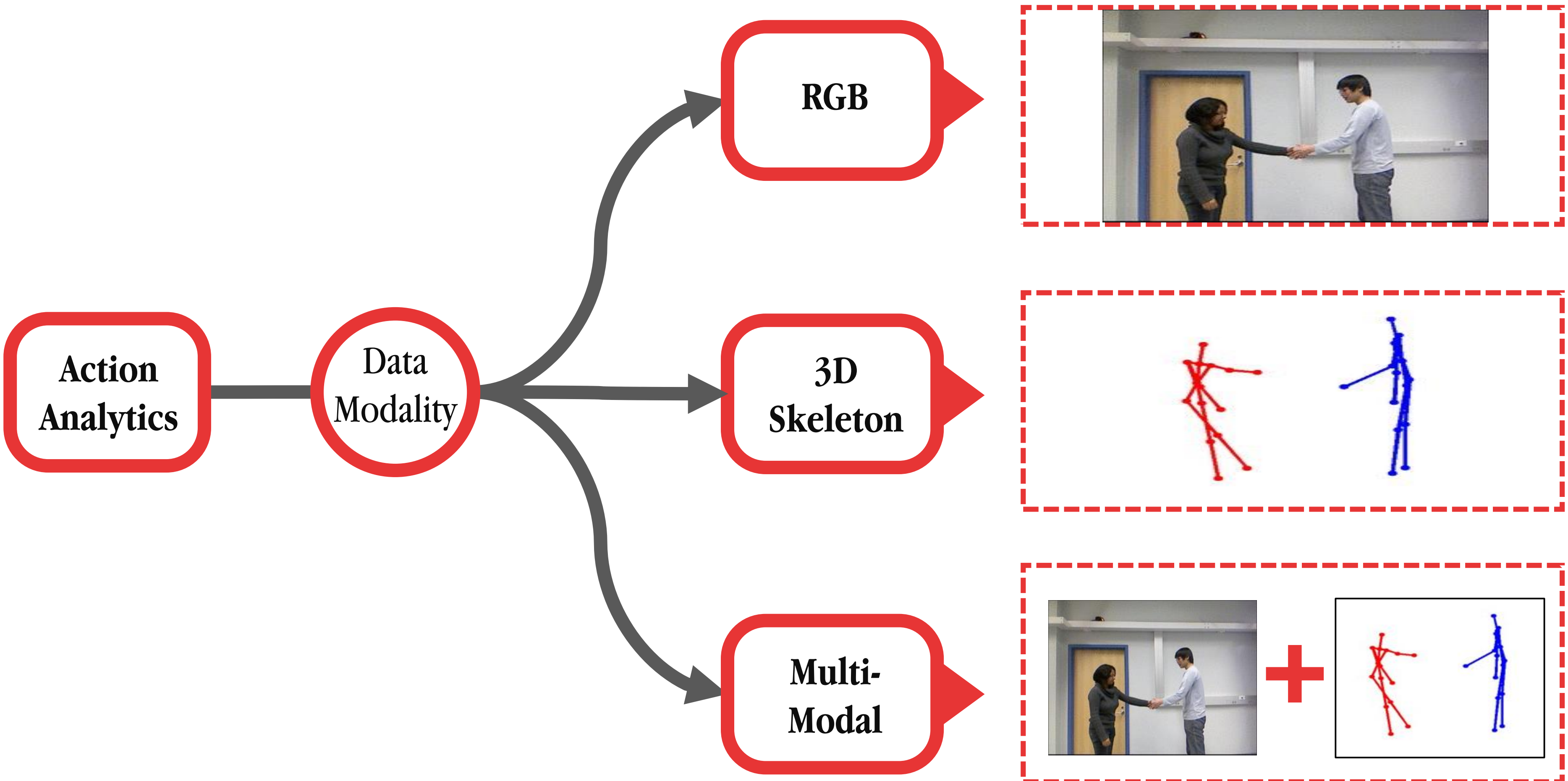
Research Methods on Action Analytics

- **Traditional methods**
  - Handcraft features + traditional classifiers
    - Limited in abilities of representation and discrimination

- **Deep learning based methods**
  - Powerful modeling ability + Large scale of video data

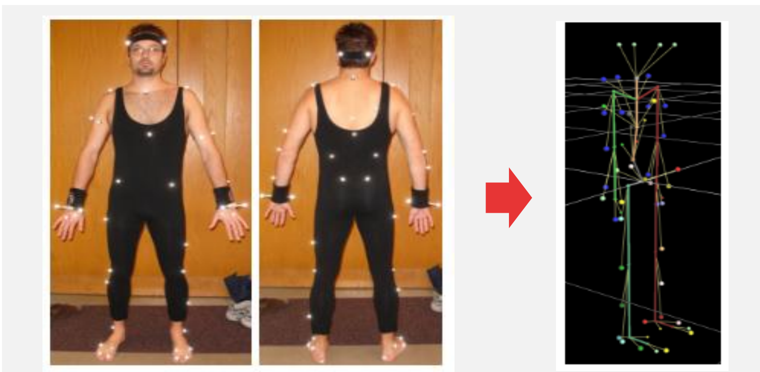
→ Current state-of-the-art action analysis methods

Data Modality on Action Analytics

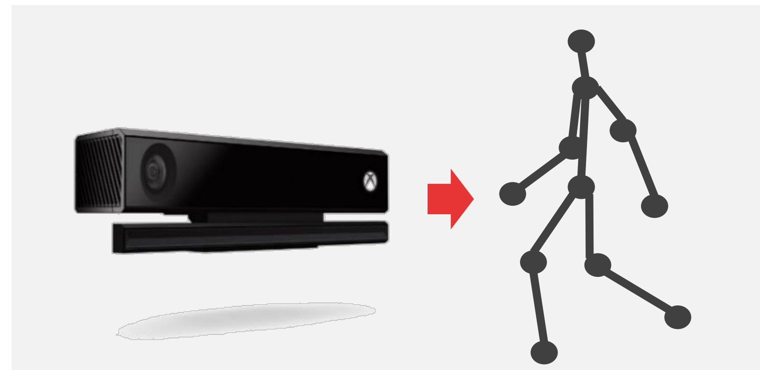


Skeleton Data

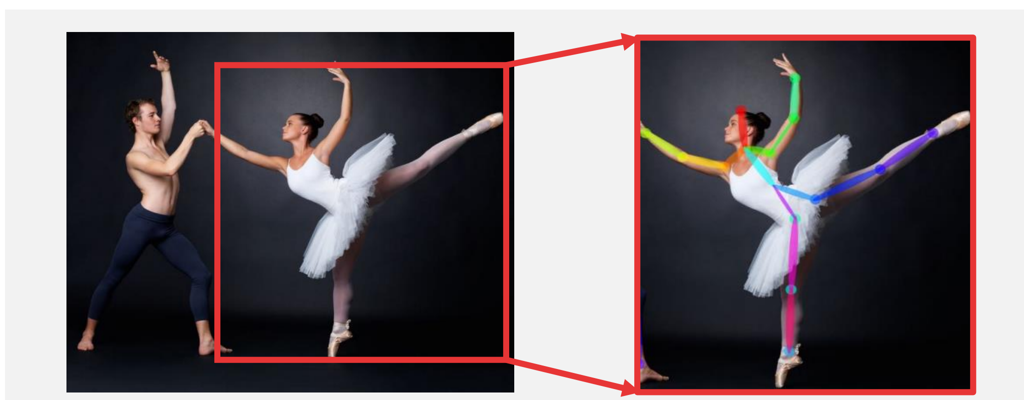
Data Access



Motion Capture System



Microsoft Kinect



Pose Estimation

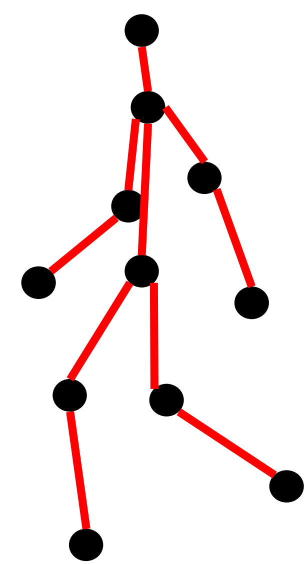
Skeleton Data

Pros

- High-level human representation
- Robust to illumination and clustered background
- Additional depth information
- Real-time online performance

Cons

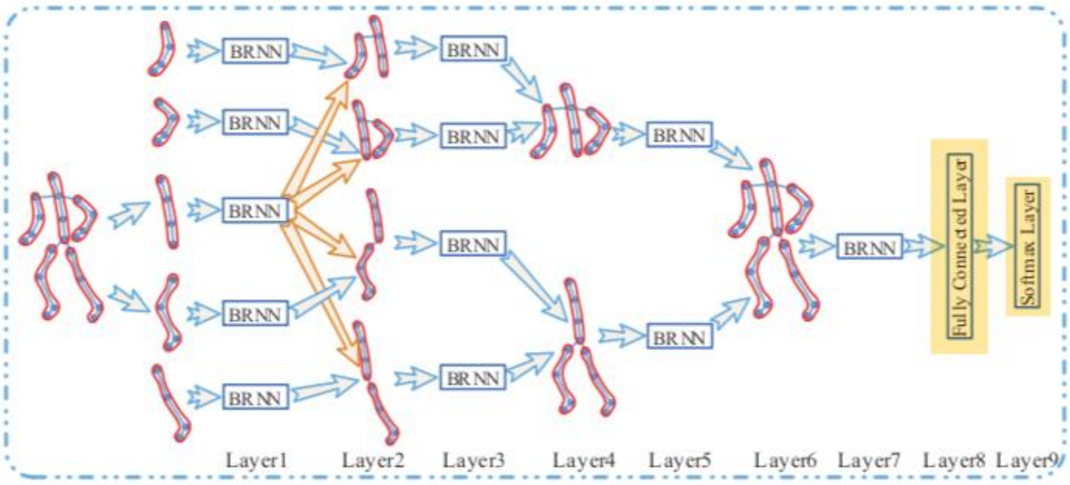
- Missing visual information
- Not reliable due to noise and occlusion



Representative Work

2015 CVPR
Hierarchical RNN

- Hierarchical Network Structure
  - Divide skeleton into five parts
- Use LSTM as Recurrent Layer
  - Long-term modeling



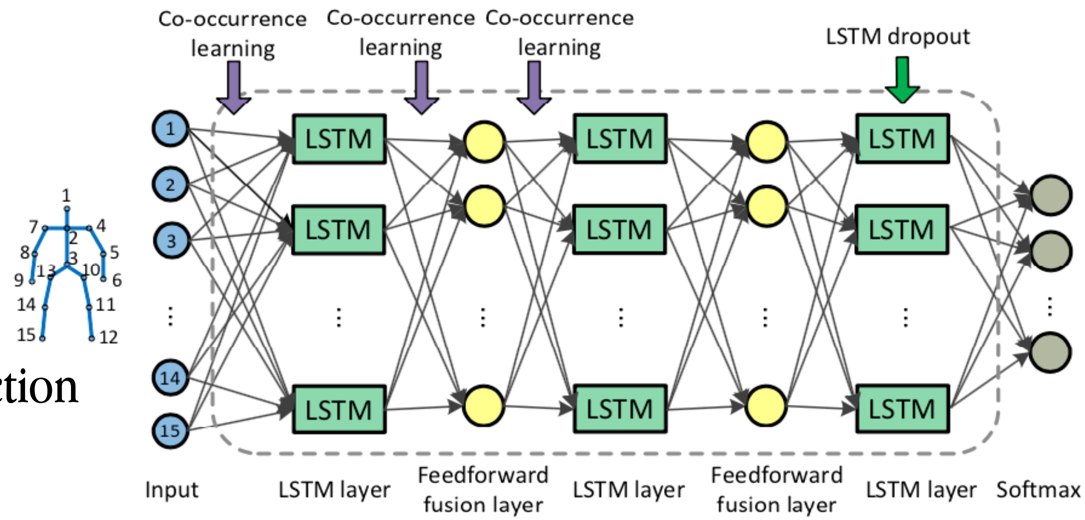
Yong Du, Wei Wang, Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition, CVPR, 2015.

Representative Work

2015 CVPR
Hierarchical RNN

2016 AAAI
Co-occurrence

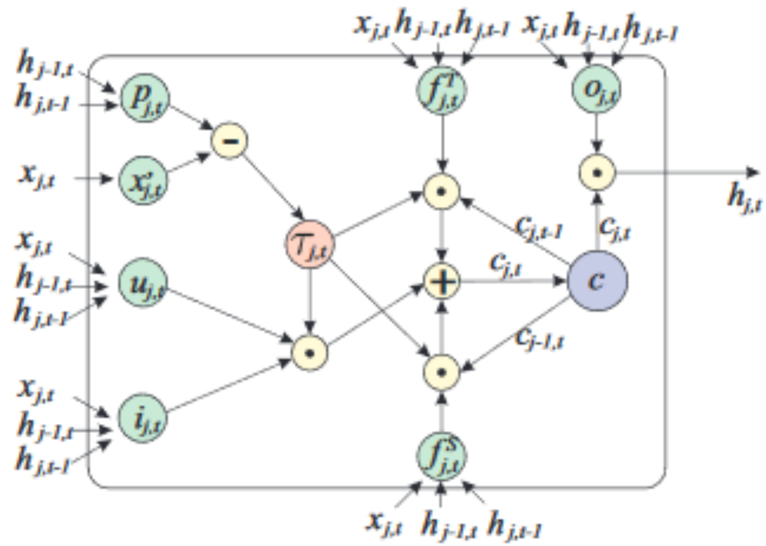
- Deep LSTM Network
  - 3 LSTM layers
  - 2 feedforward layers
- Co-occurrence Feature Learning
  - Regularization term in loss function
- In-depth Dropout for LSTM



Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks , AAAI, 2016.

Representative Work

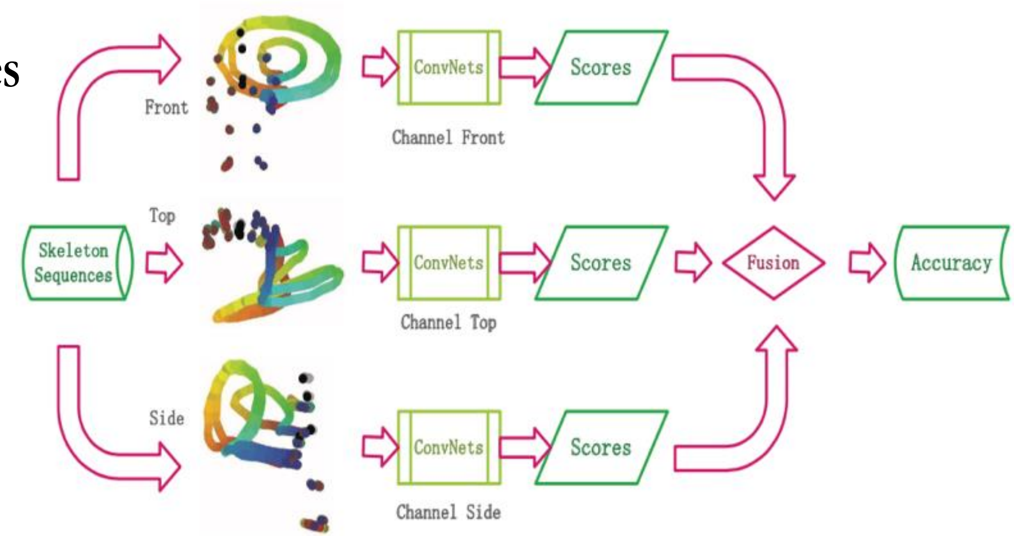
- Spatio-Temporal LSTM with Trust Gates
  - Tree Structure based Traversal
    - Keep the kinematic dependency
  - Trust Gates
    - Analyze reliability of each input
    - Exclude noise and occlusion



Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, Alex C Kot. Global context-aware attention LSTM networks for 3D action recognition, CVPR, 2017

Representative Work

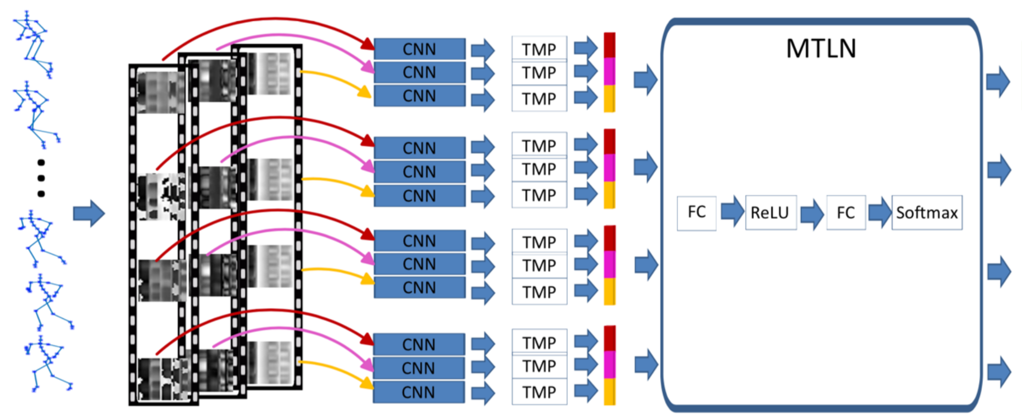
- Encode skeleton sequence → 2D images
  - 3 Joint Trajectory Maps (JTM)
  - 3 ConvNets
- Joint Trajectory Maps
  - Encoding joint motion direction
  - Encoding body parts
  - Encoding motion magnitude



Pichao Wang, Zhaoyang Li, Yonghong Hou, Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks, MM, 2016

Representative Work

- A New representation
  - Three clips with cylindrical coordinates
- Deep CNNs
  - Extract hierarchical features
- Multi-Task learning network
  - Jointly process features in parallel

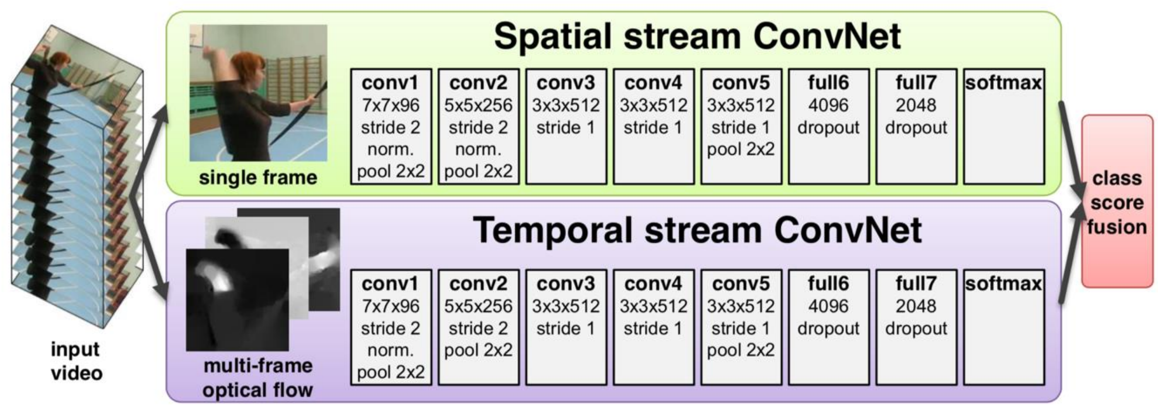


Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, Farid Boussaid. A new representation of skeleton sequences for 3d action recognition, CVPR, 2017

Representative Work

2014 NIPS
Two-Stream

- Spatial Stream
  - Single RGB frame
- Temporal Stream
  - Stacked optical flow frames
- Multi-Task Learning
  - UCF101 + HMDB51



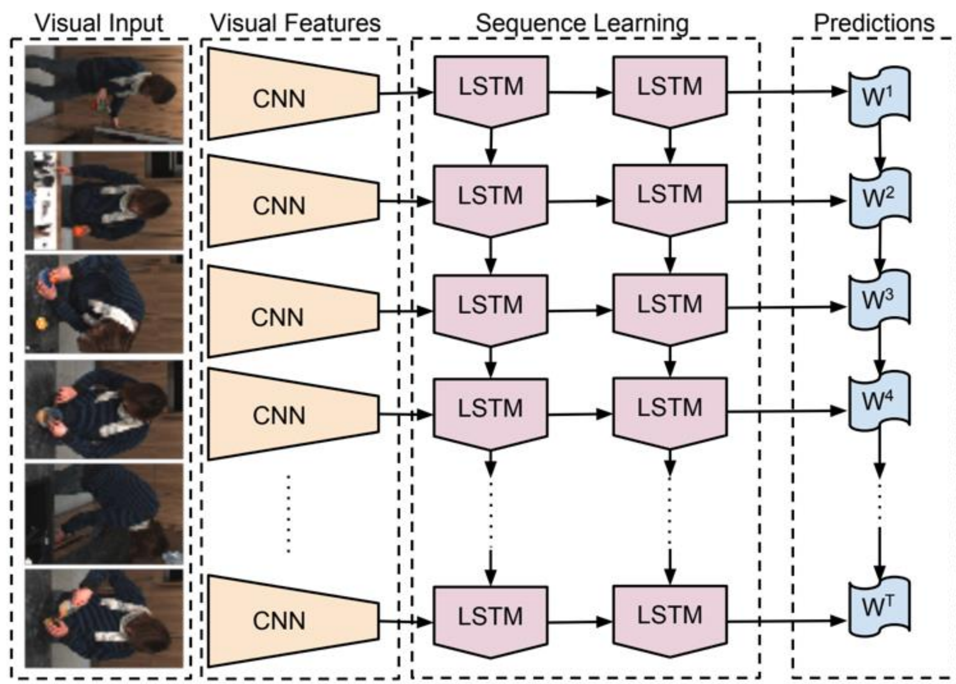
Karen Simonyan, Andrew Zisserman. Two-stream convolutional networks for action recognition in videos, NIPS, 2014

Representative Work

2014 NIPS
Two-Stream

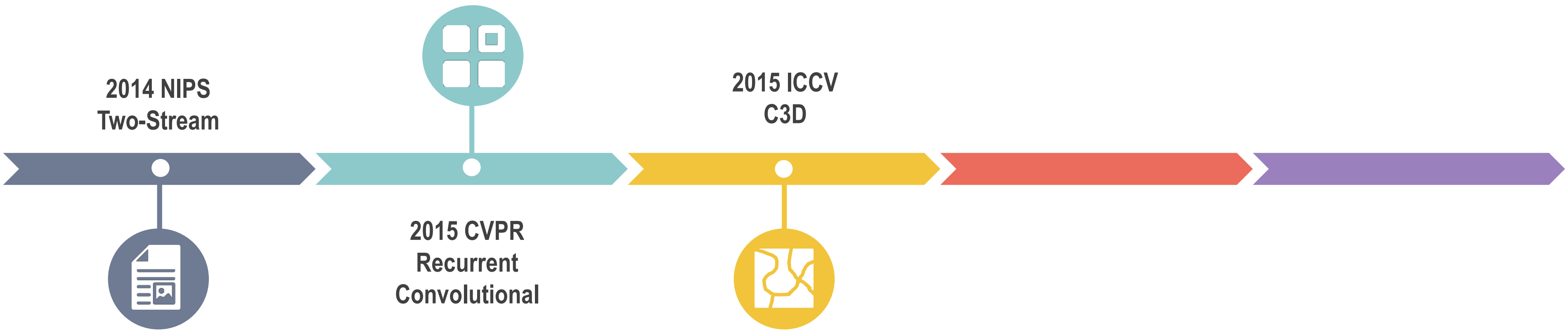
2015 CVPR
Recurrent
Convolutional

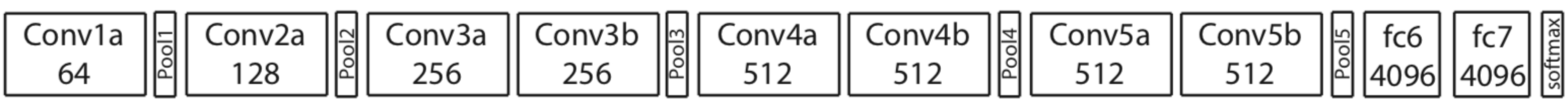
- CNN
  - Visual features for single frame
- LSTM
  - Long-term temporal modeling
- Different applications
  - Action recognition
  - Image/Video description



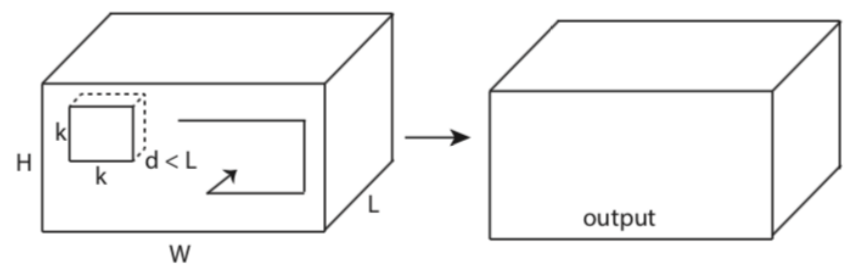
Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell. Long-Term recurrent convolutional networks for visual recognition and description, CVPR, 2015

Representative Work



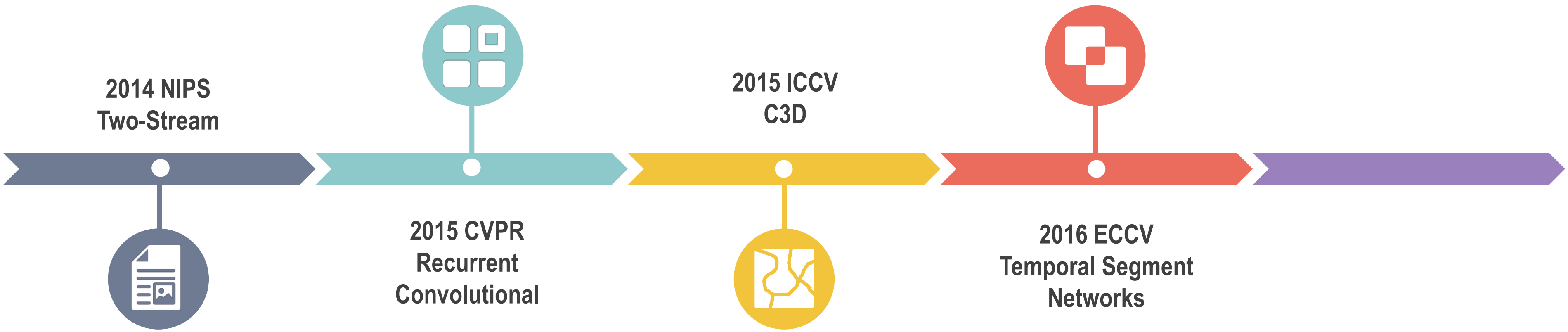


- Convolutional 3D
  - 3D convolution + 3D pooling
  - 8 convolution + 5 pooling
- 3x3x3 kernel works best

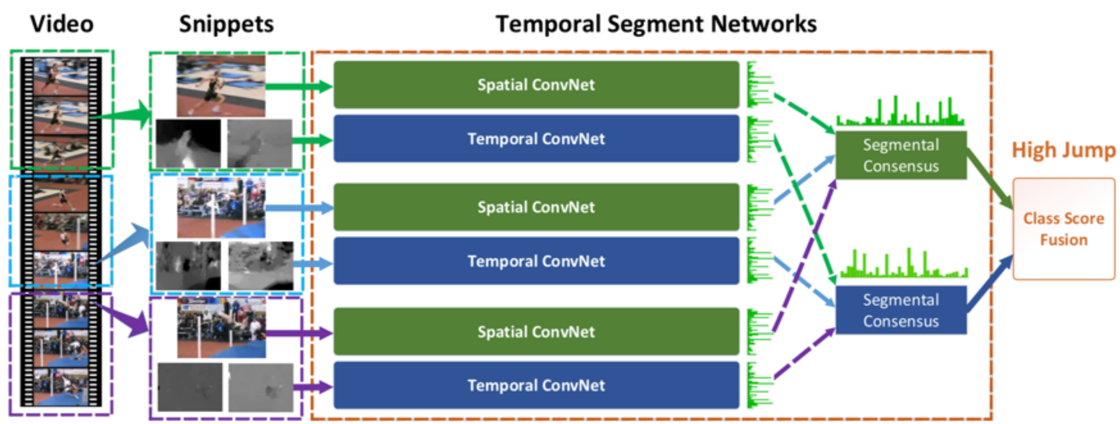


Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks, ICCV, 2015

Representative Work

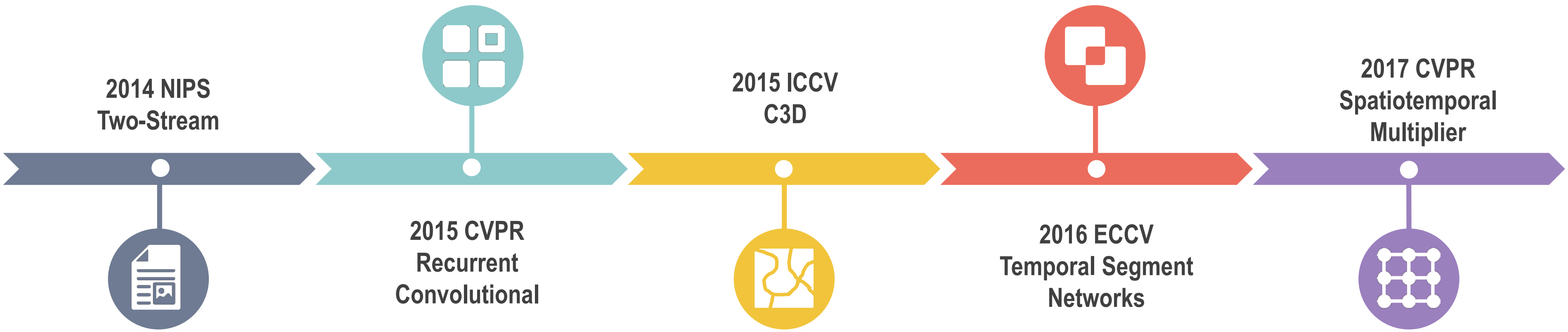


- Temporal Segment Networks
  - Sparse temporal sampling
  - Video-level supervision
- Good practices on video data
  - Cross-modality pre-training
  - Data augmentation
  - Partial BN

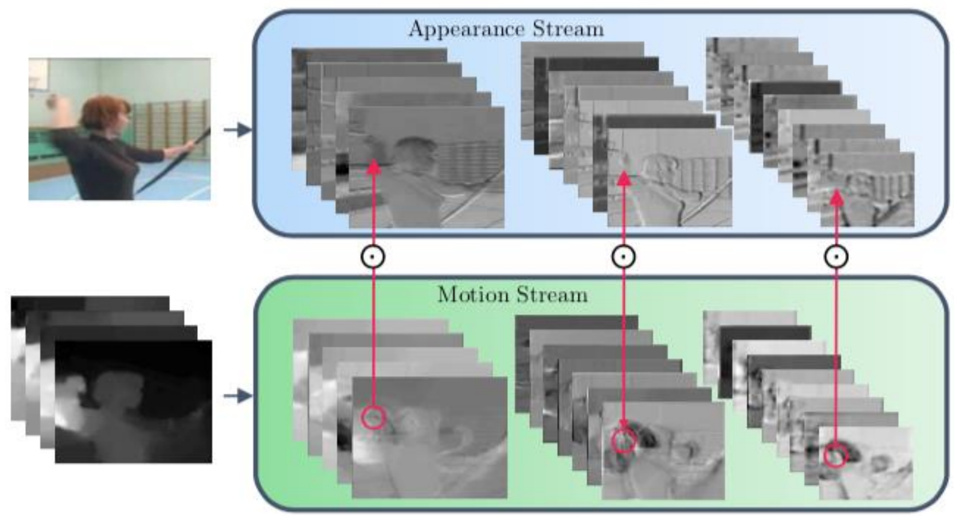


Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, Luc Van Gool. Temporal segment networks: towards good practices for deep action recognition, ECCV, 2016

Representative Work



- Spatio-Temporal Multiplier Network
  - Two-stream structure
  - Multiplicative interaction
- Temporal filtering
  - 1D temporal convolutions

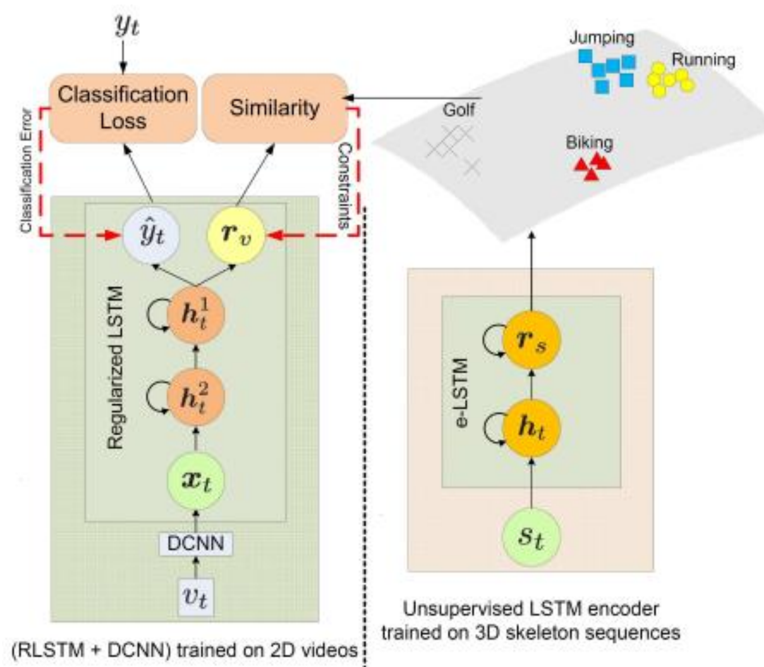


Christoph Feichtenhofer, Axel Pinz, Richard Wildes. Spatiotemporal Multiplier Networks for Video Action Recognition, CVPR, 2017

Representative Work

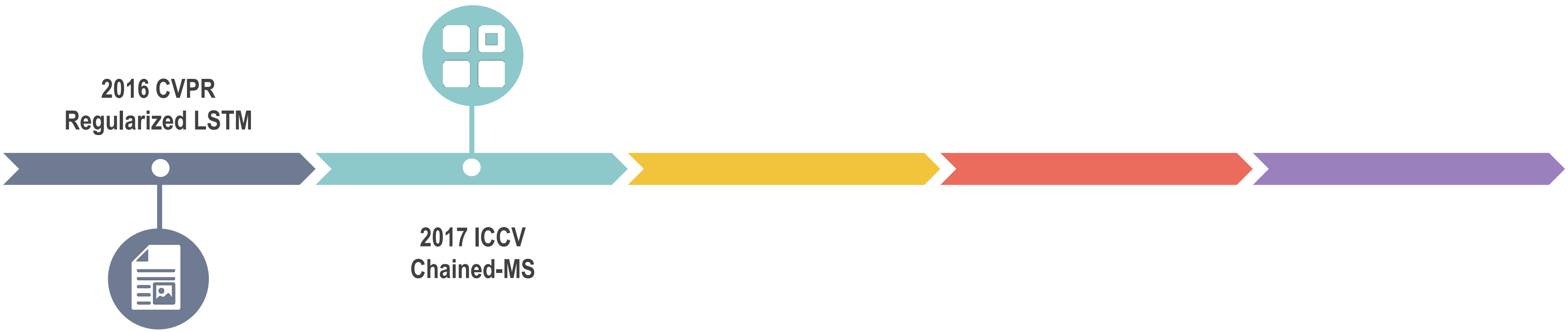
2016 CVPR
Regularized LSTM

- Explore the common space
  - RGB: RLSTM+DCNN
  - Skeleton: Unsupervised LSTM
- Regularization on long-term modeling
  - Class independent regularization
  - Class Specific Regularization

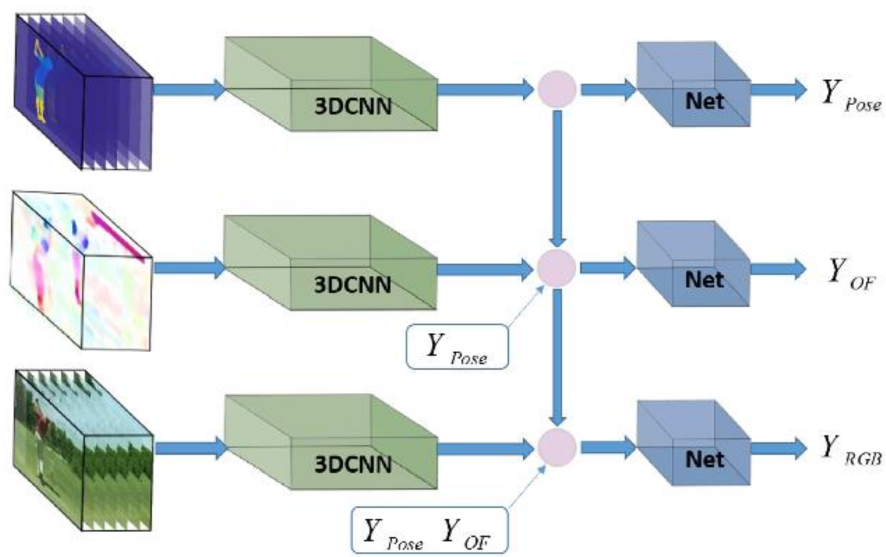


Mahasseni, Behrooz, and Sinisa Todorovic. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. CVPR 2016.

Representative Work



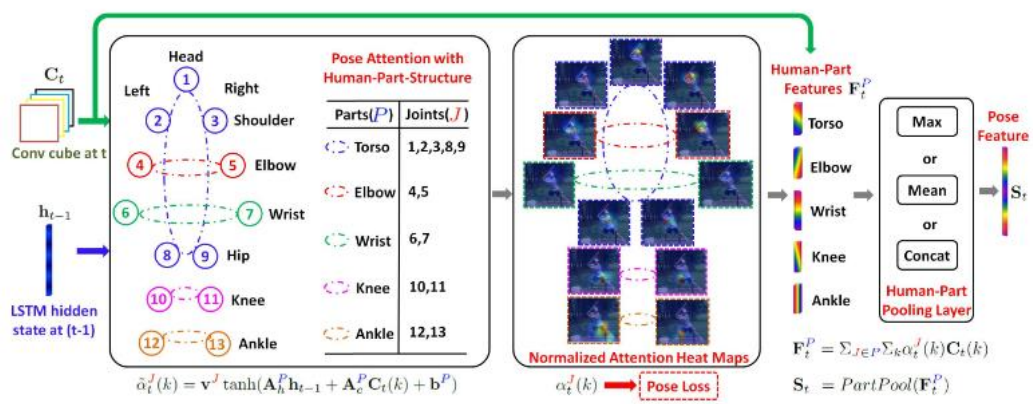
- Chained multi-stream network
  - RGB
  - Optical flow
  - Body part segmentation
- Multi-stream fusion
  - Markov chain



Zolfaghari, Mohammadreza, et al. "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection." ICCV 2017.

Representative Work

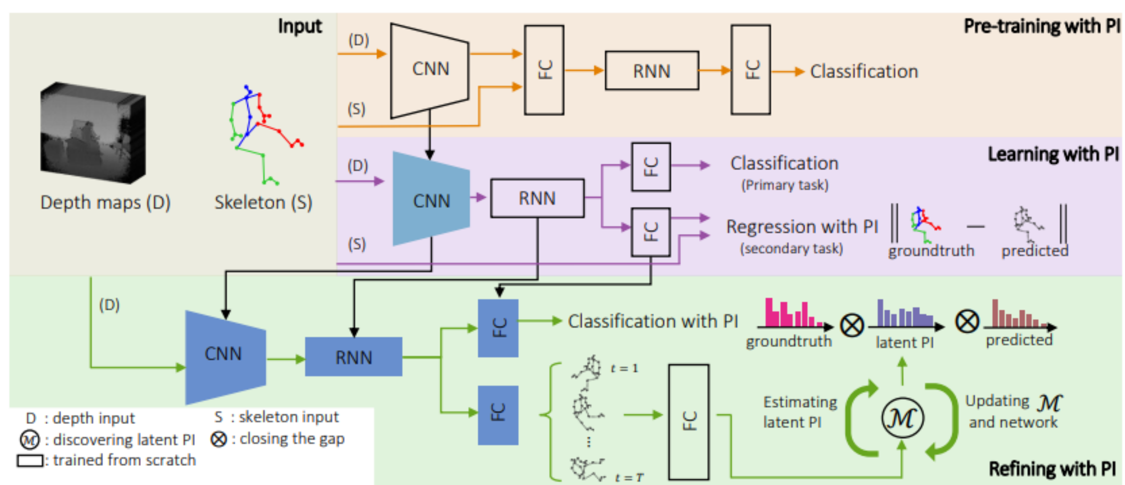
- Multi-task network
  - Pose estimation
  - Action recognition
- Pose attention
  - Indicate human structure



Wenbin Du, Yali Wang, and Yu Qiao. "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos." ICCV 2017

Representative Work

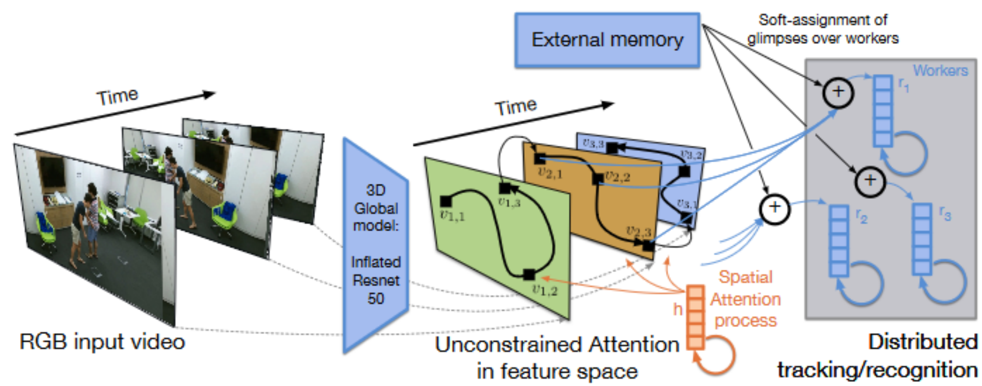
- Privileged information
  - Regression with PI
  - Classification with PI
- PI-RNN
  - Pre-train with PI
  - Learn with PI
  - Refine with PI



Shi, Zhiyuan, and Tae-Kyun Kim. "Learning and refining of privileged information-based rnns for action recognition from depth sequences." CVPR 2017

Representative Work

- Make glimpse as humans
  - Recurrent attention model
  - Distributed soft-tracking workers
  - External Memory



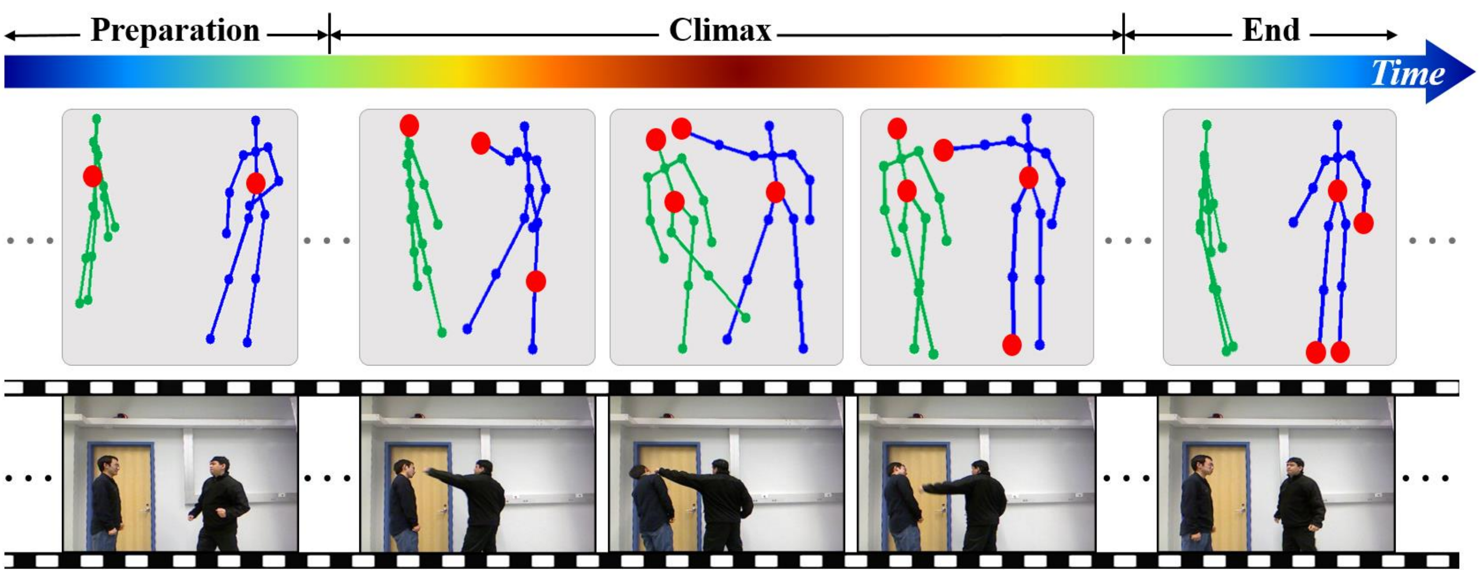
Baradel, Fabien, et al. "Glimpse clouds: Human activity recognition from unstructured feature points." CVPR 2018.

Attention Based Action Recognition

Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng and Jiaying Liu. "An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data", *AAAI* 2017.

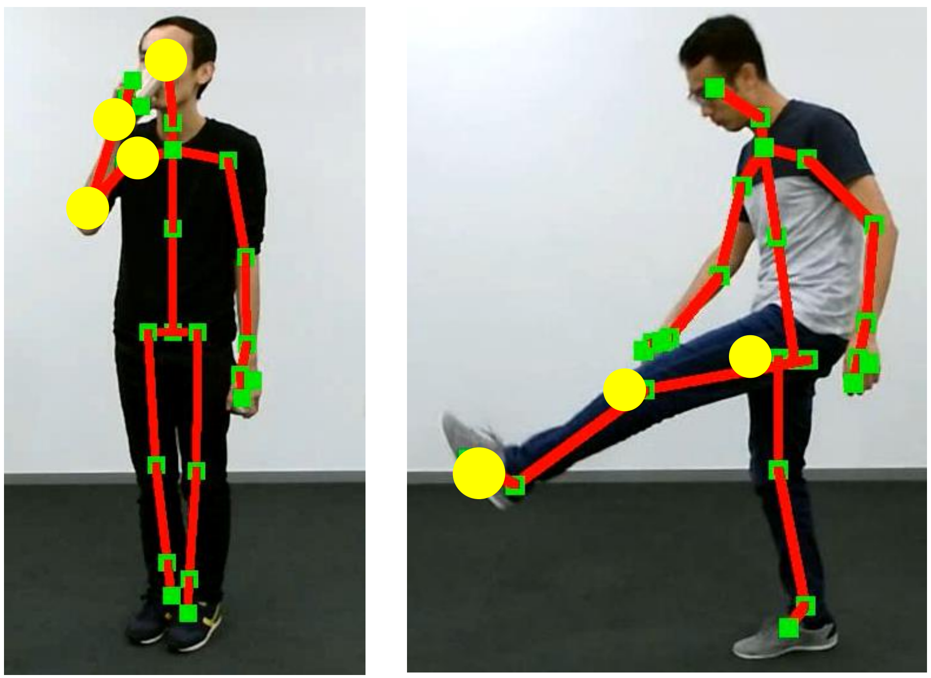
Spatio-Temporal Attention Model for Action Recognition

- Spatial attention to select key joints
- Temporal attention to extract key frames

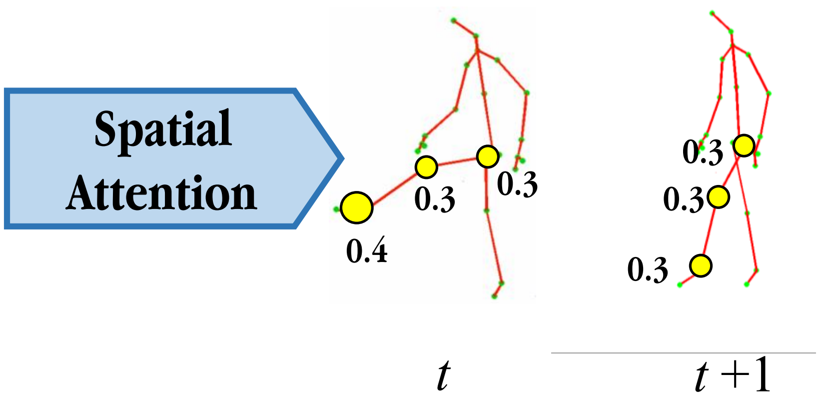


Spatial Attention

- **Importance of joints varies**
  - Learn spatial attention model
  - Add different weights to joints

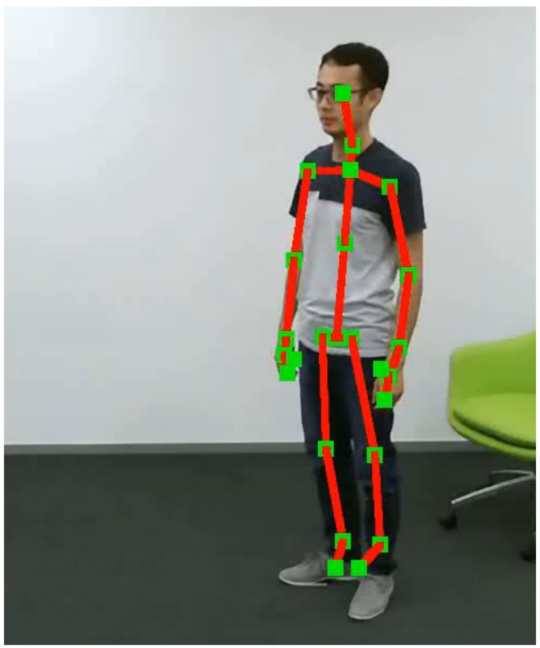


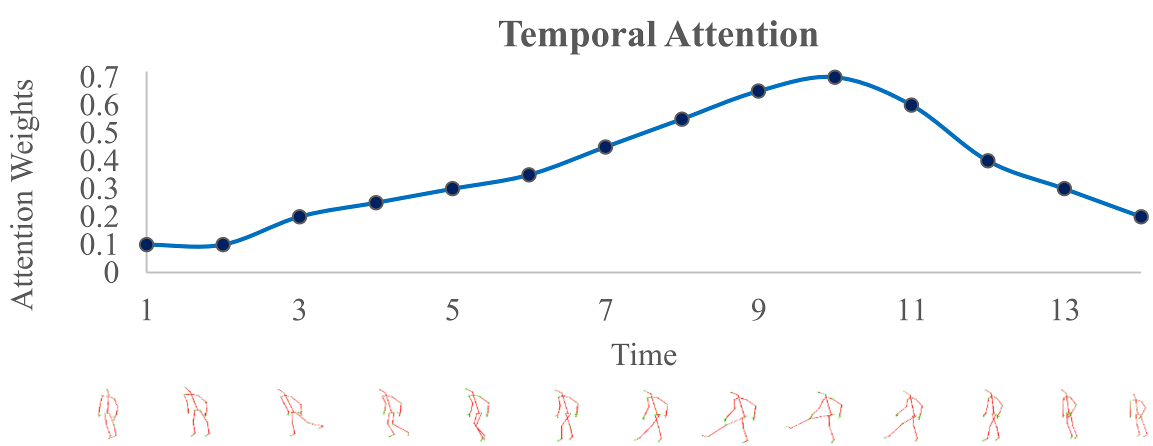
Drinking   Kicking



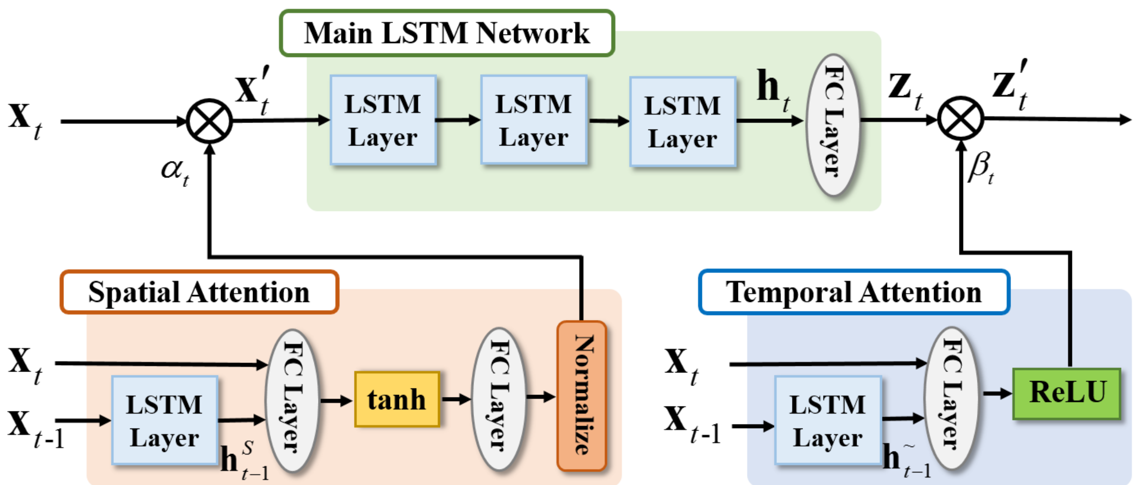
Temporal Attention

- **Importance of frames varies**
  - Assign attention weights to each frame
  - Fuse the output of all frames based on the attention





Spatio-Temporal Attention



Spatial Attention:

Allocate different weights to each joint automatically.

$$\mathbf{s}_t = U_s \tanh(W_{xs}\mathbf{x}_t + W_{hs}\mathbf{h}^s_{t-1} + \mathbf{b}_s) + \mathbf{b}_{us}$$

$$\alpha_{t,k} = \frac{\exp(s_{t,k})}{\sum_{i=1}^{K}\exp(s_{t,i})}$$

Temporal Attention:

Allocate different importance to each frame.

$$\beta_t = \mathrm{ReLU}(\mathbf{w}_{x\sim}\mathbf{x}_t + \mathbf{w}_{h\sim}\mathbf{h}^{\sim}_{t-1} + b_{\sim})$$

Objective Function

Regularization for spatial attention

- Encourage the model to focus on more spatial joints
- Avoid local optimal

Reduce Overfitting

$$L = -\sum_{i=1}^{C} y_i \log \hat{y}_i + \lambda_1 \sum_{k=1}^{K} \left(1 - \frac{\sum_{t=1}^{T} \alpha_{t,k}}{T}\right)^2 + \frac{\lambda_2}{T} \sum_{t=1}^{T} \|\beta_t\|_2 + \lambda_3 \|W_{uv}\|_1$$

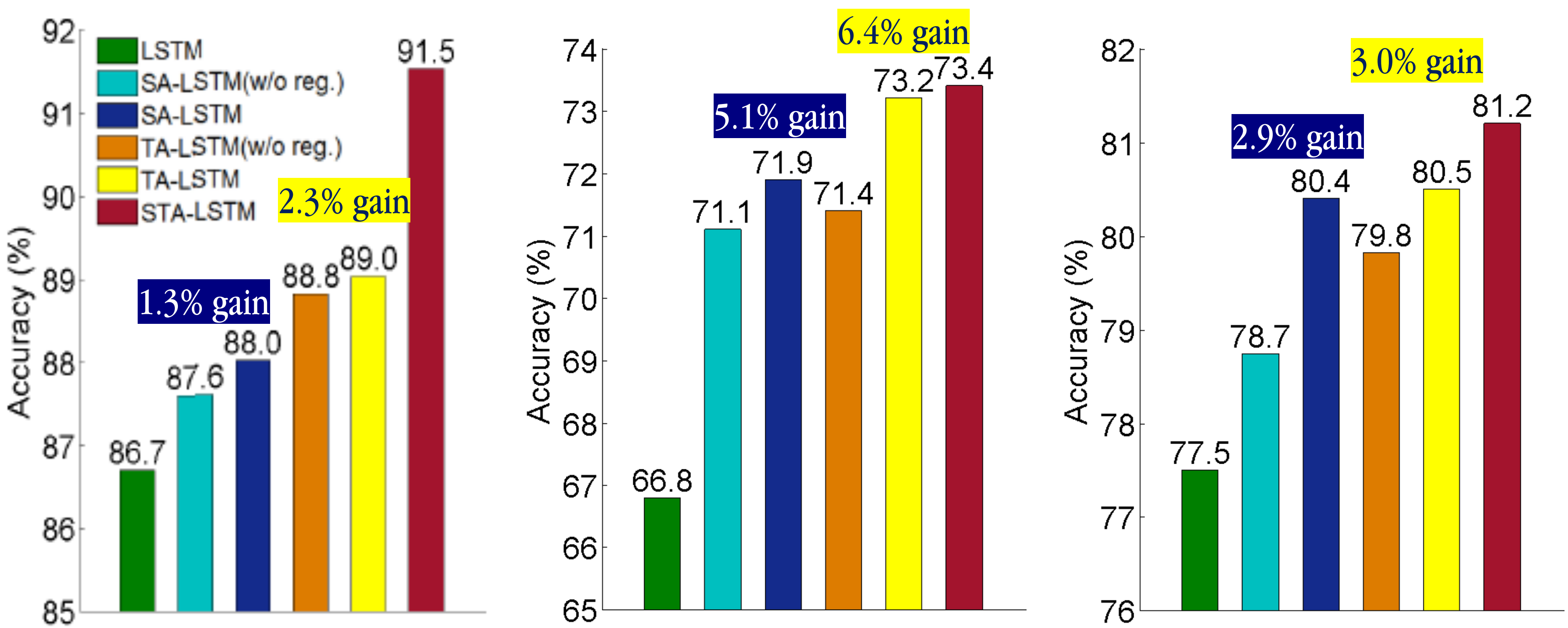
Cross Entropy

Regularization for temporal attention

- Avoid temporal attention values increasing unboundedly
- Avoid gradient vanishing

Experimental Results

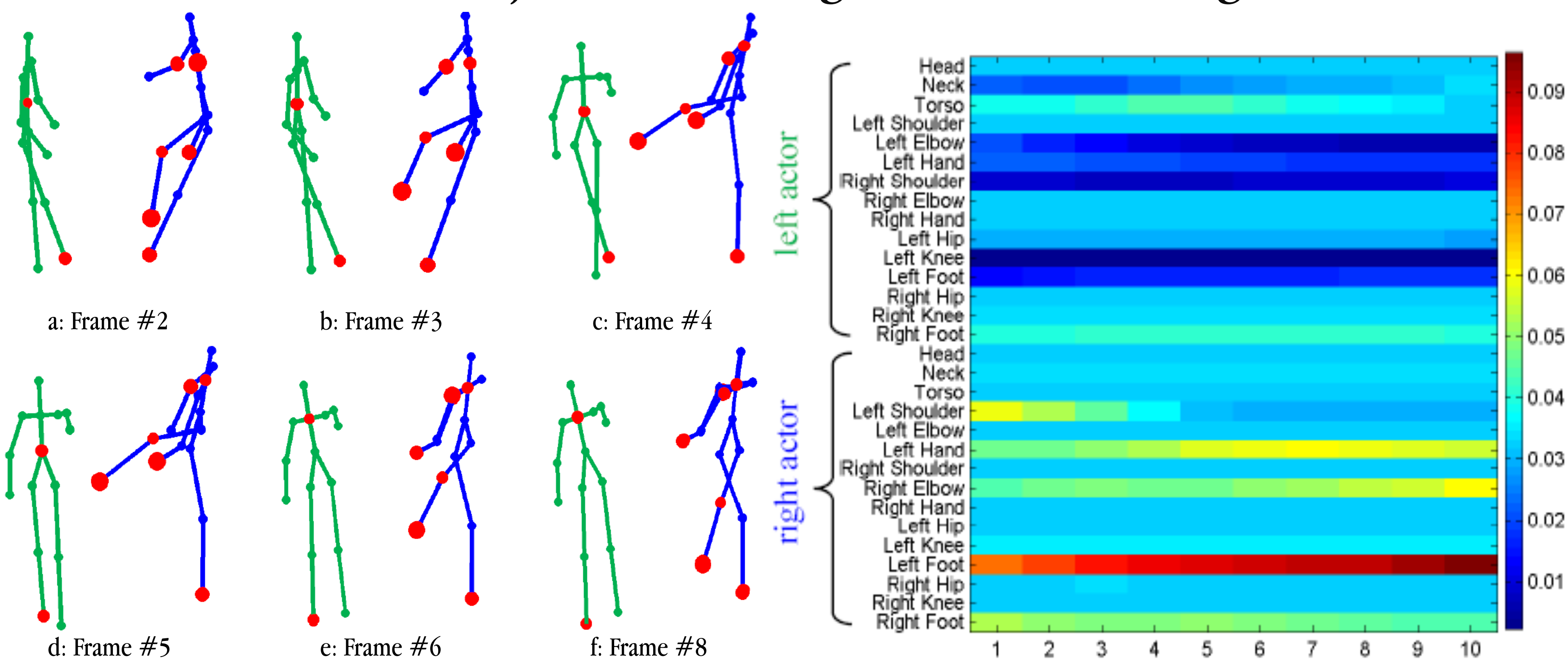
- Action recognition results
  - SBU Kinect Interaction Dataset (SBU)
  - NTU RGB+D Dataset (NTU)



(a) SBU (b) NTU-Cross Subject (c) NTU-Cross View

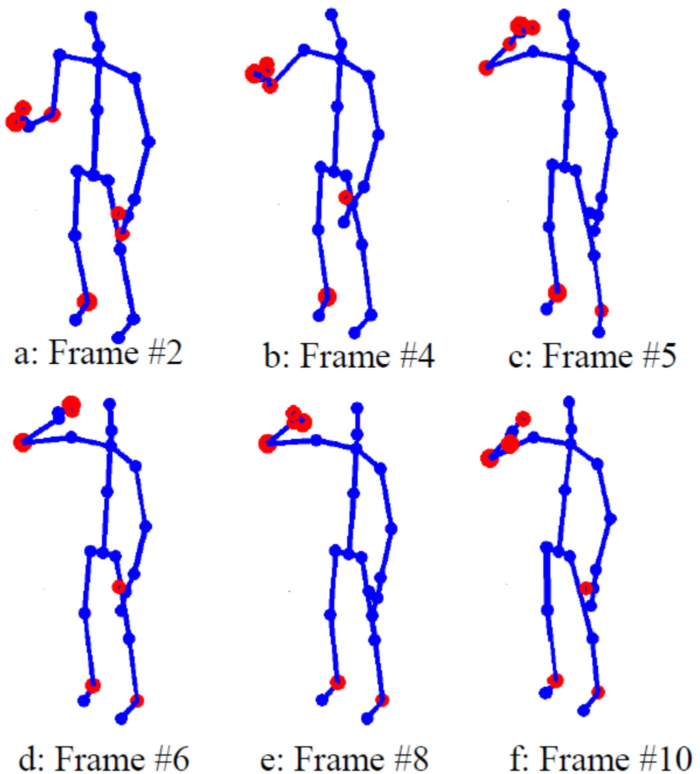
Experimental Results

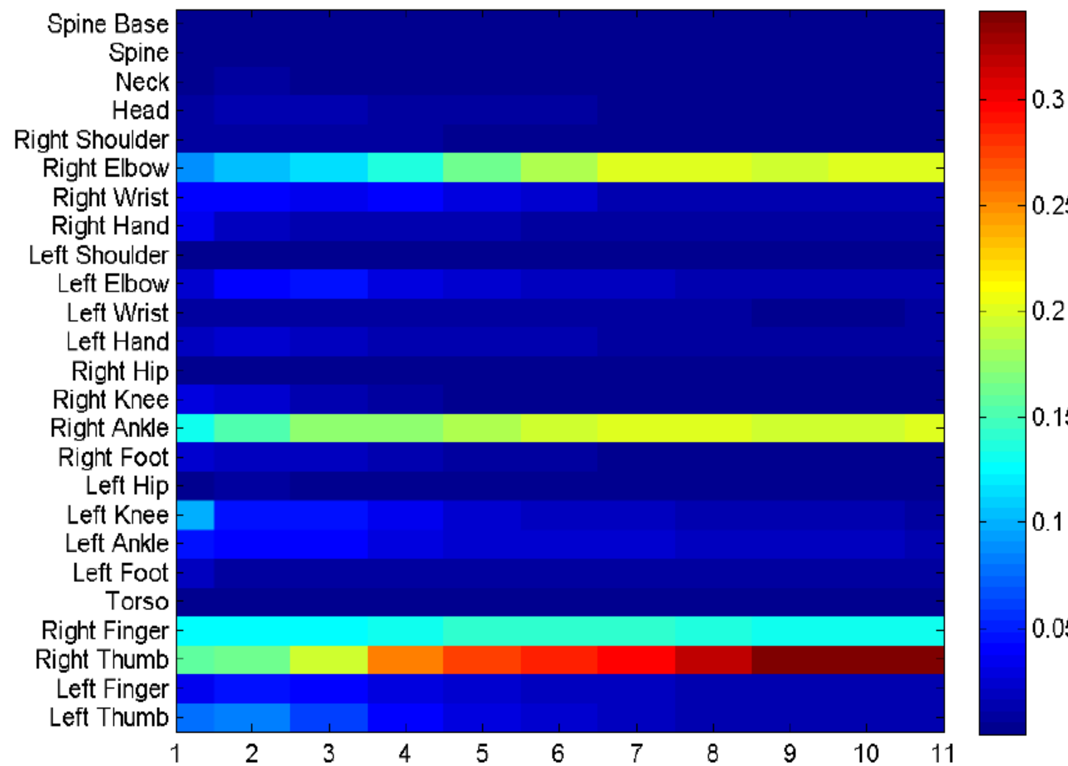
- Visualization of Spatial Attention
  - Action 'Kicking'
  - Red circle: 8 joints with largest attention weights



Experimental Results

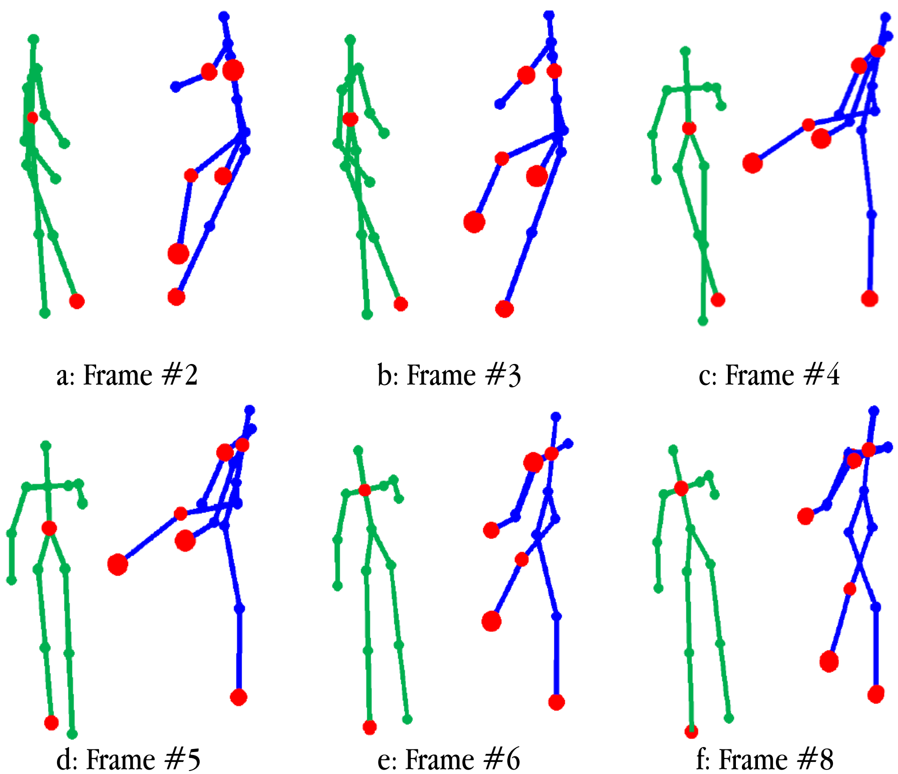
- Visualization of Spatial Attention
  - Action ‘Drinking’
  - Red circle: 8 joints with largest attention weights

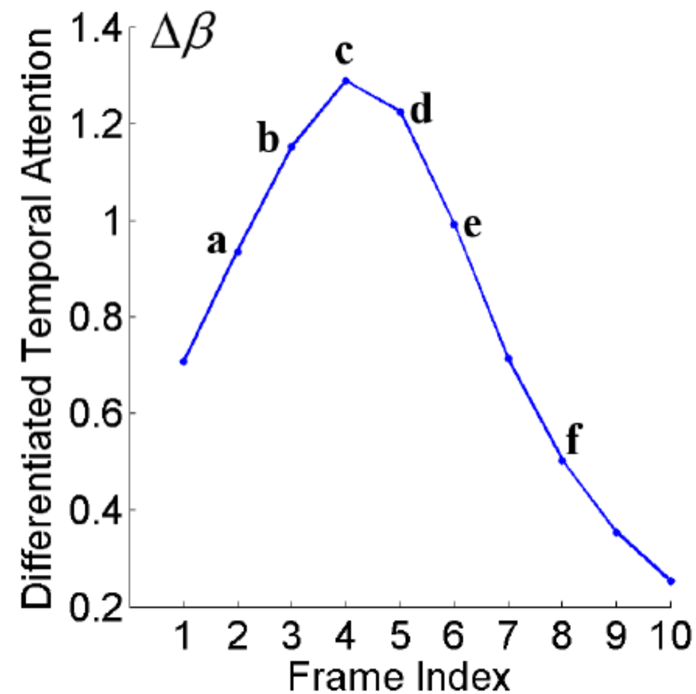




Experimental Results

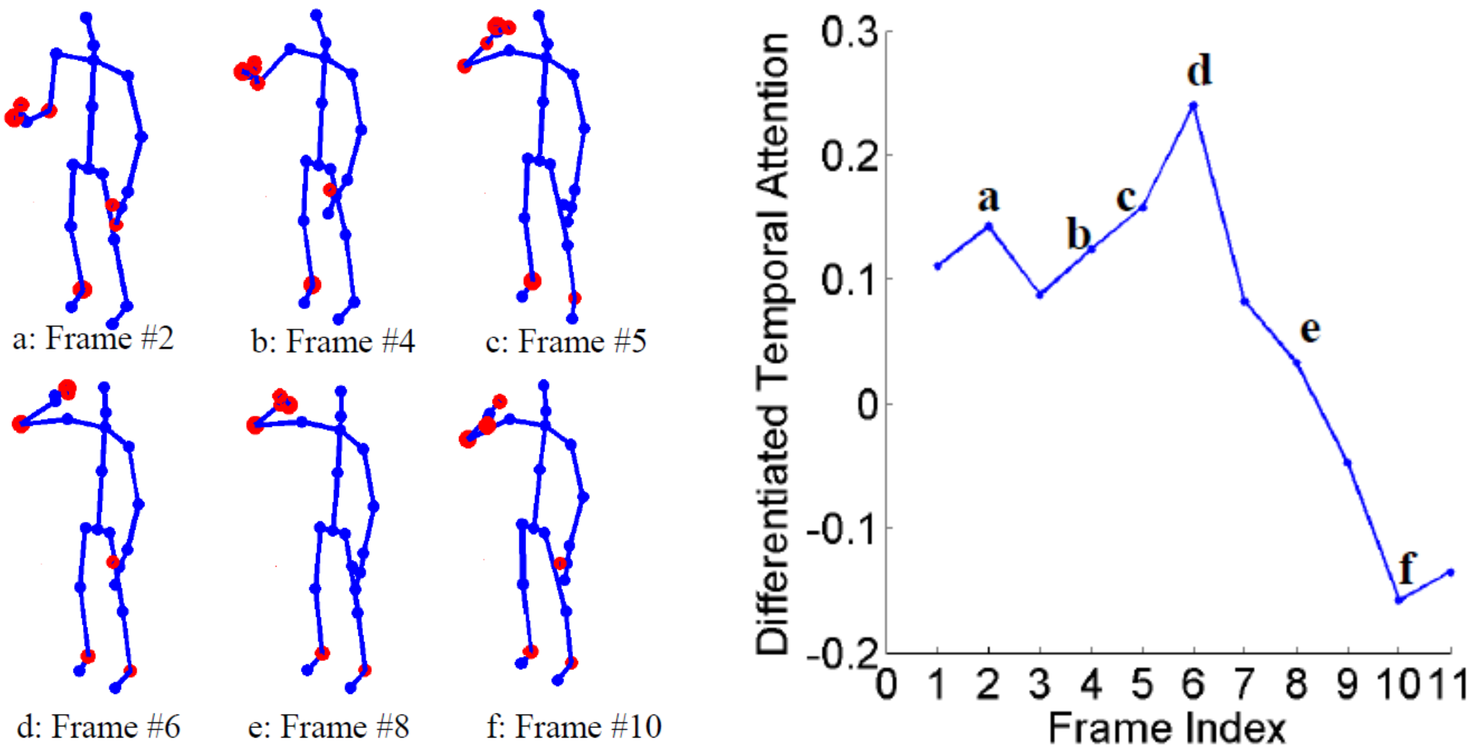
- Visualization of Temporal Attention
  - Action 'Kicking'





Experimental Results

- Visualization of Temporal Attention
  - Action 'Drinking'

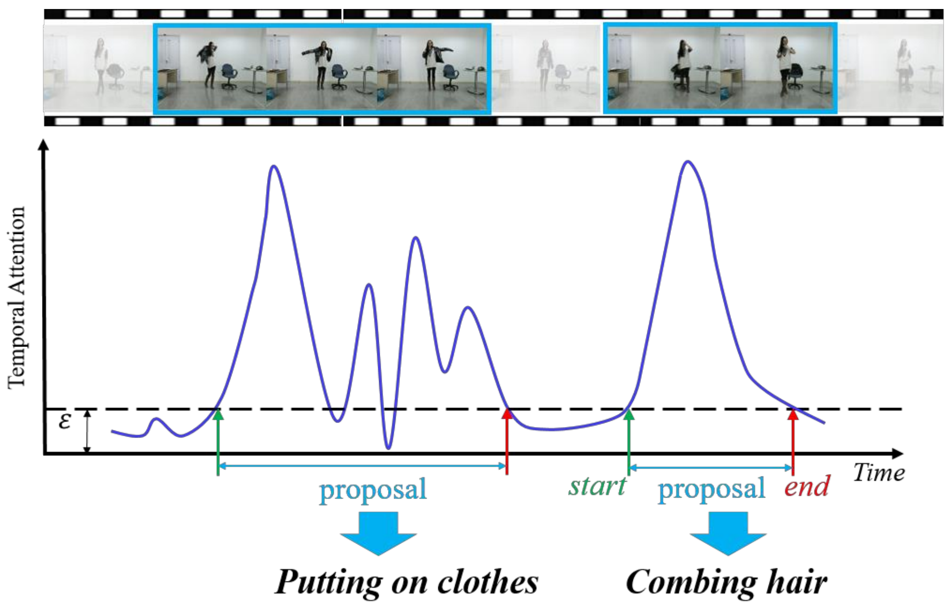


Attention based Action Detection

Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng and Jiaying Liu. "Spatial-Temporal Attention Based LSTM Networks for 3D Action Recognition and Detection", *IEEE Trans. on Image Processing* (*TIP*), Vol.27, No.7, pp.3459-3471, July 2018.

Temporal Attention Based Action Detection

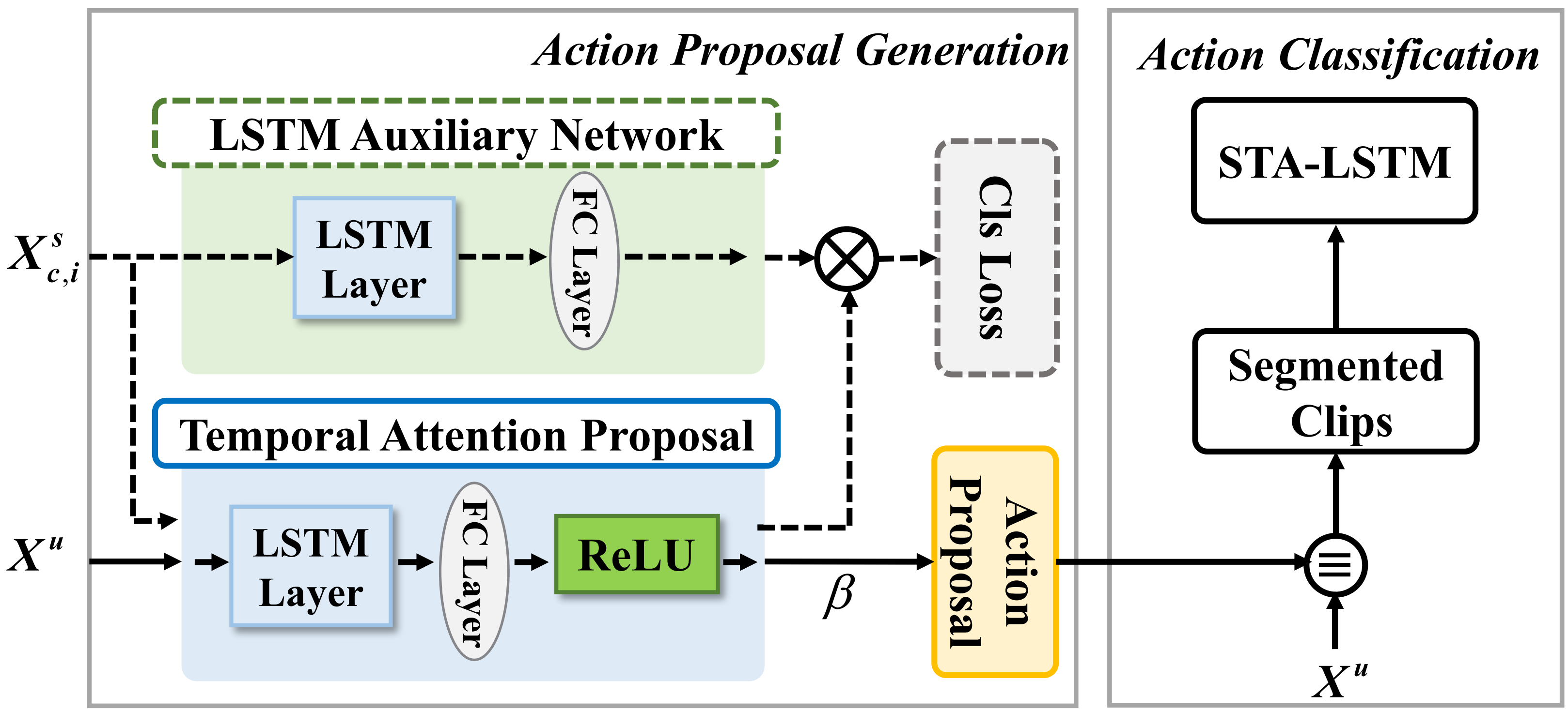
- Temporal Response for Action Localization
- Action Classification



Action Detection

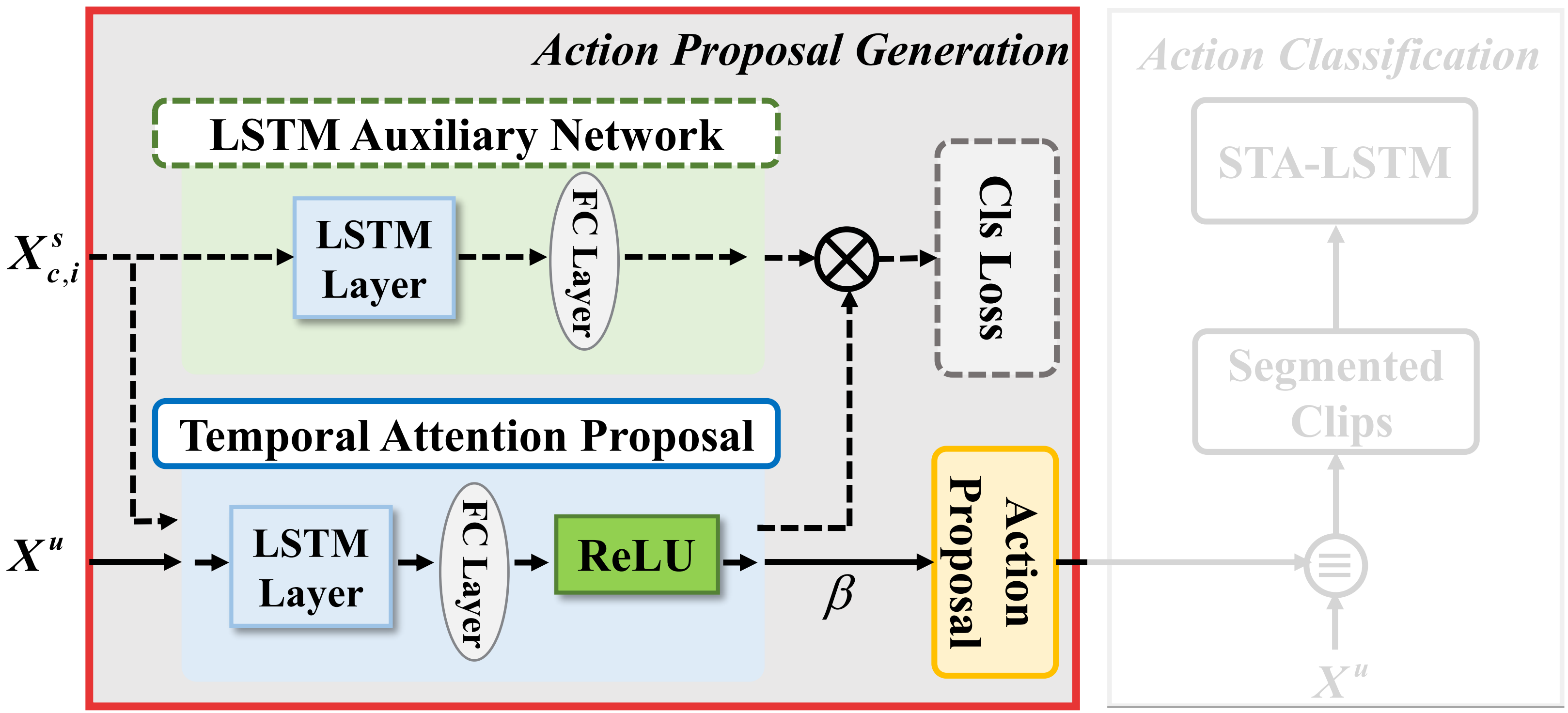
Network Structure

- Action Proposal Generation
- Action Classification



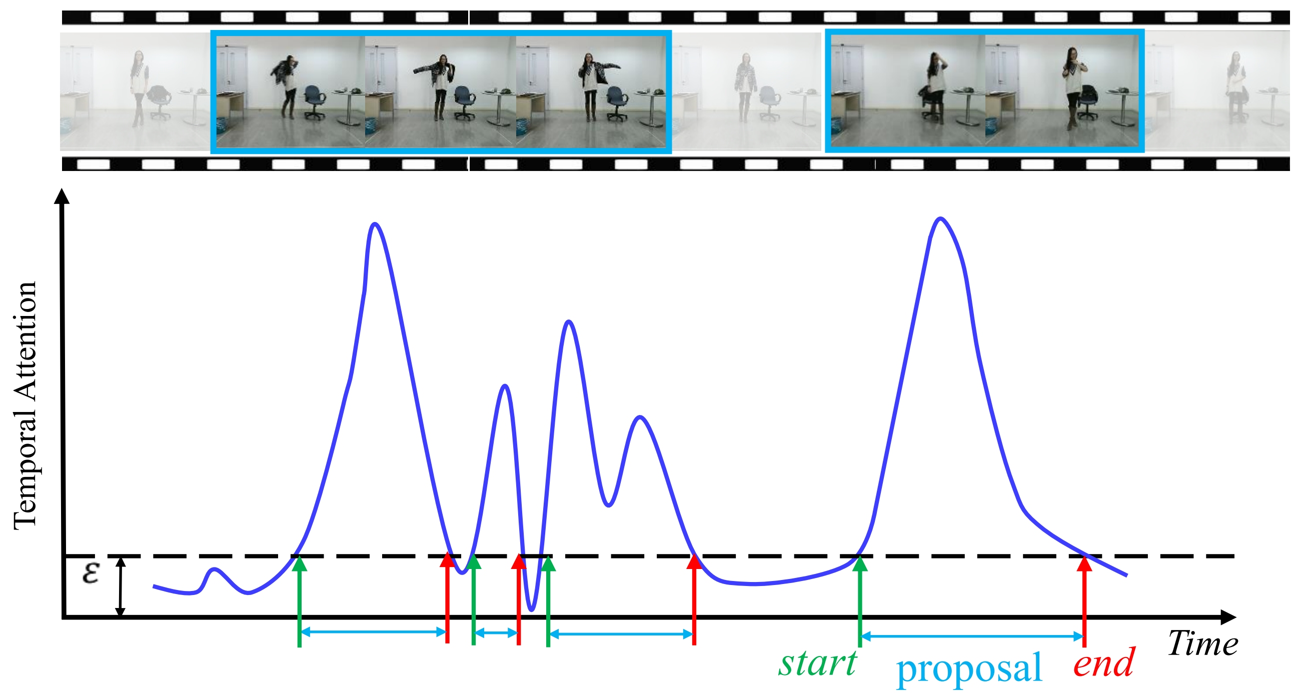
Action Proposal Generation

- **Temporal Attention Proposal**
  - Pretrain with Trimmed Video
  - Temporal Attention → Temporal Response



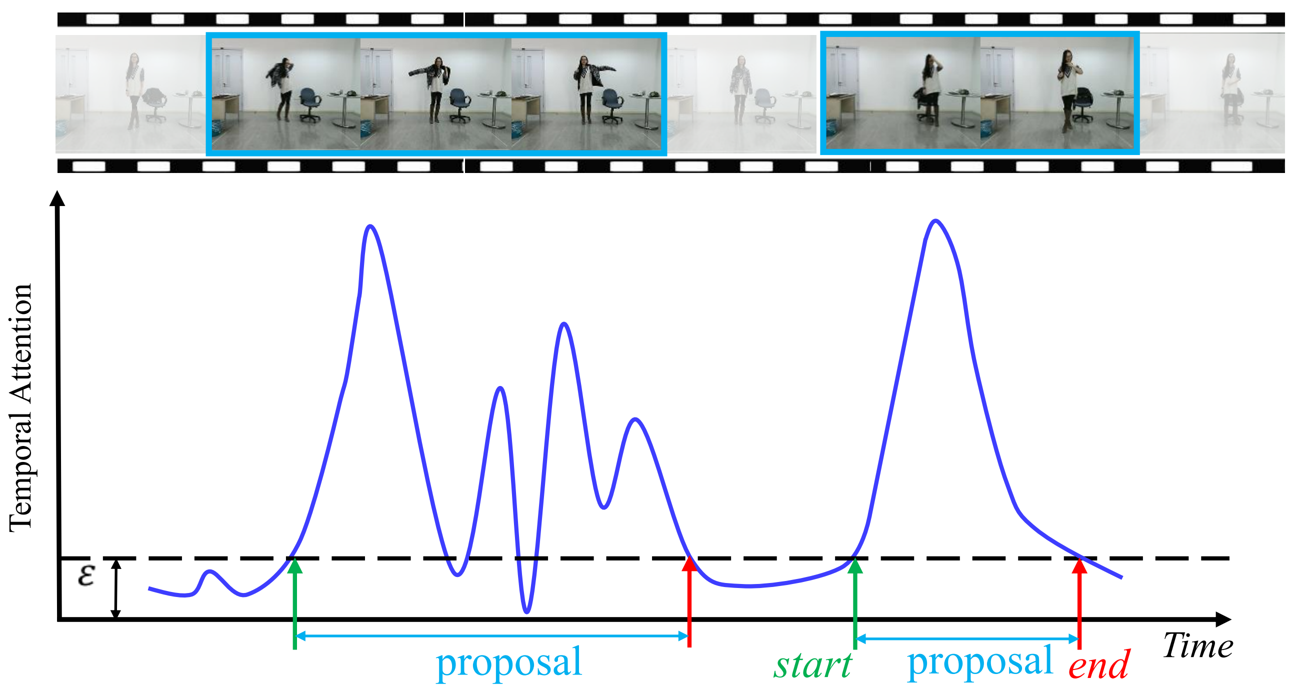
Action Proposal Generation

- **Temporal Attention Proposal**
  - Temporal Response > ε → Action Proposal



Action Proposal Generation

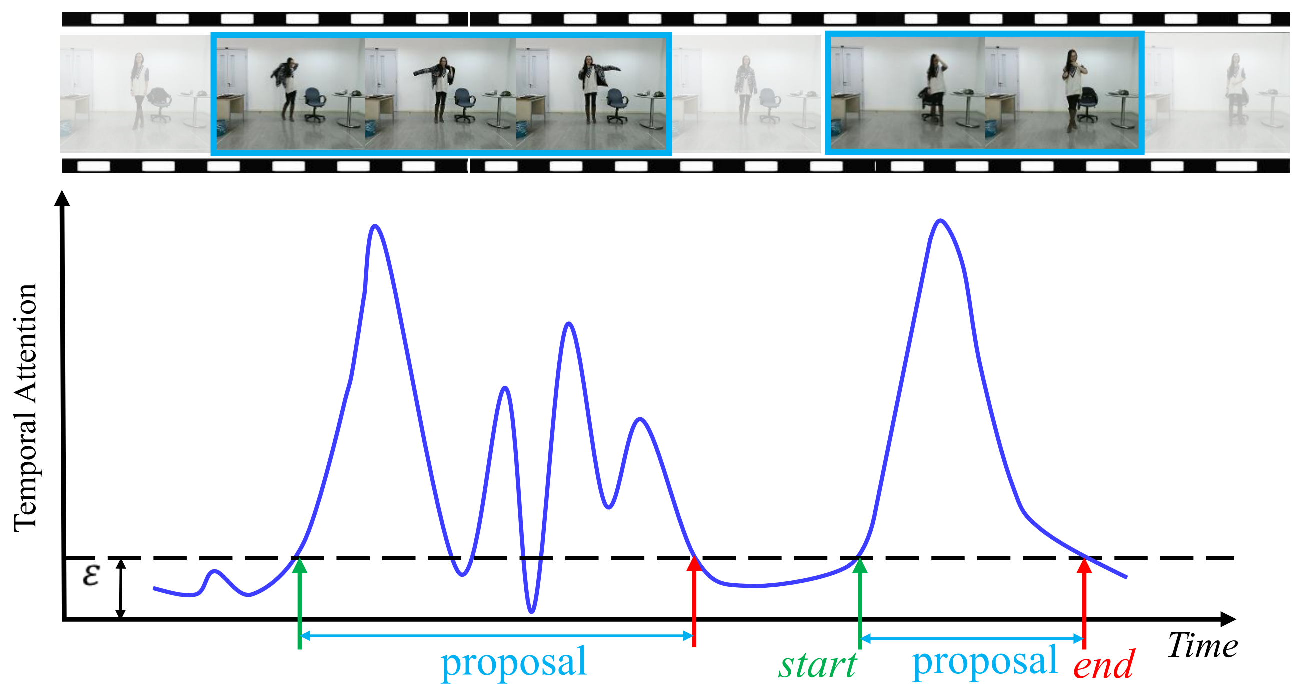
- **Temporal Attention Proposal**
  - Temporal Response > ε → Action Proposal
  - Adjacent Proposals Aggregation → Less Fragments



Action Proposal Generation

Temporal Attention Proposal

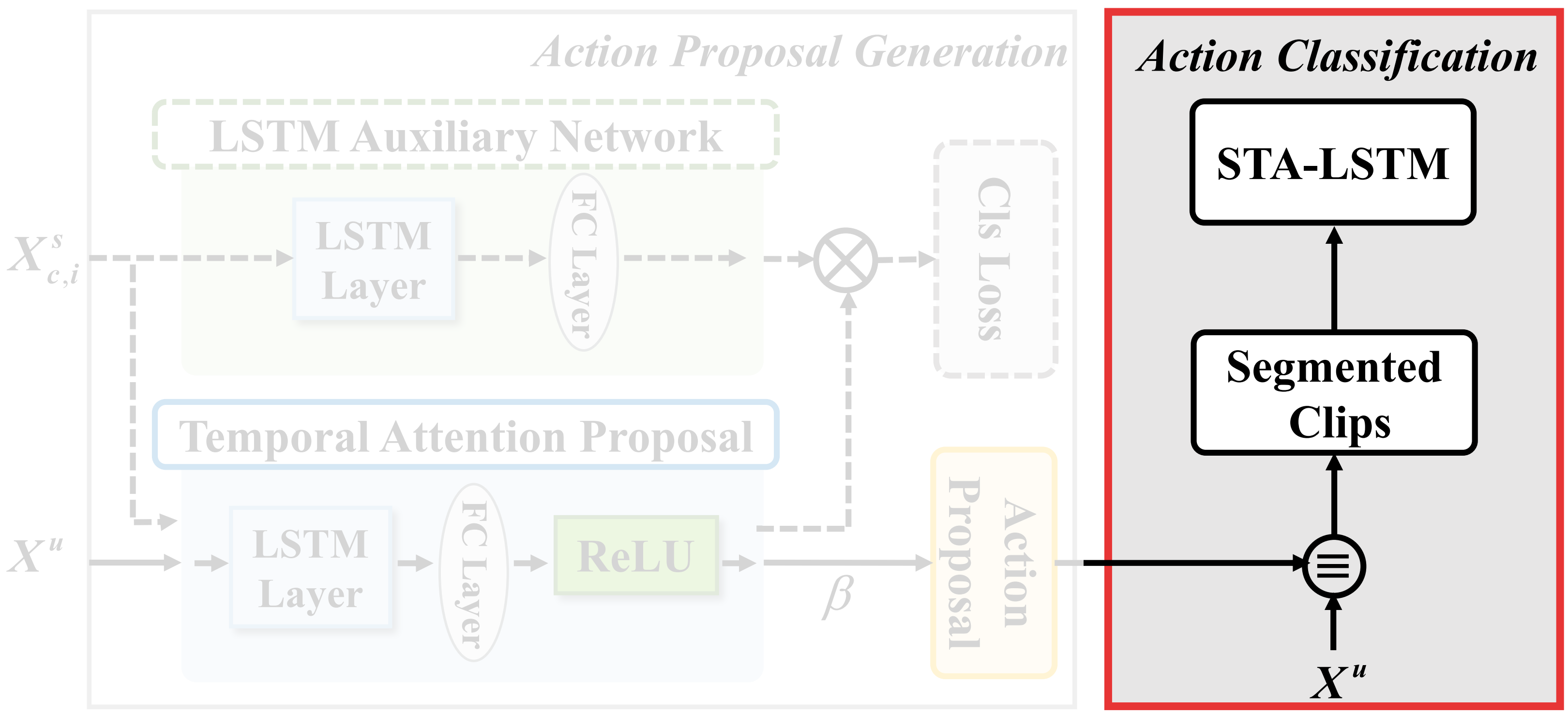
- Temporal Response > ε → Action Proposal
- Adjacent Proposals Aggregation → Less Fragments
- Multiscale Scheme → More Accurate Localization



Action Classification

Classification on Each Proposal

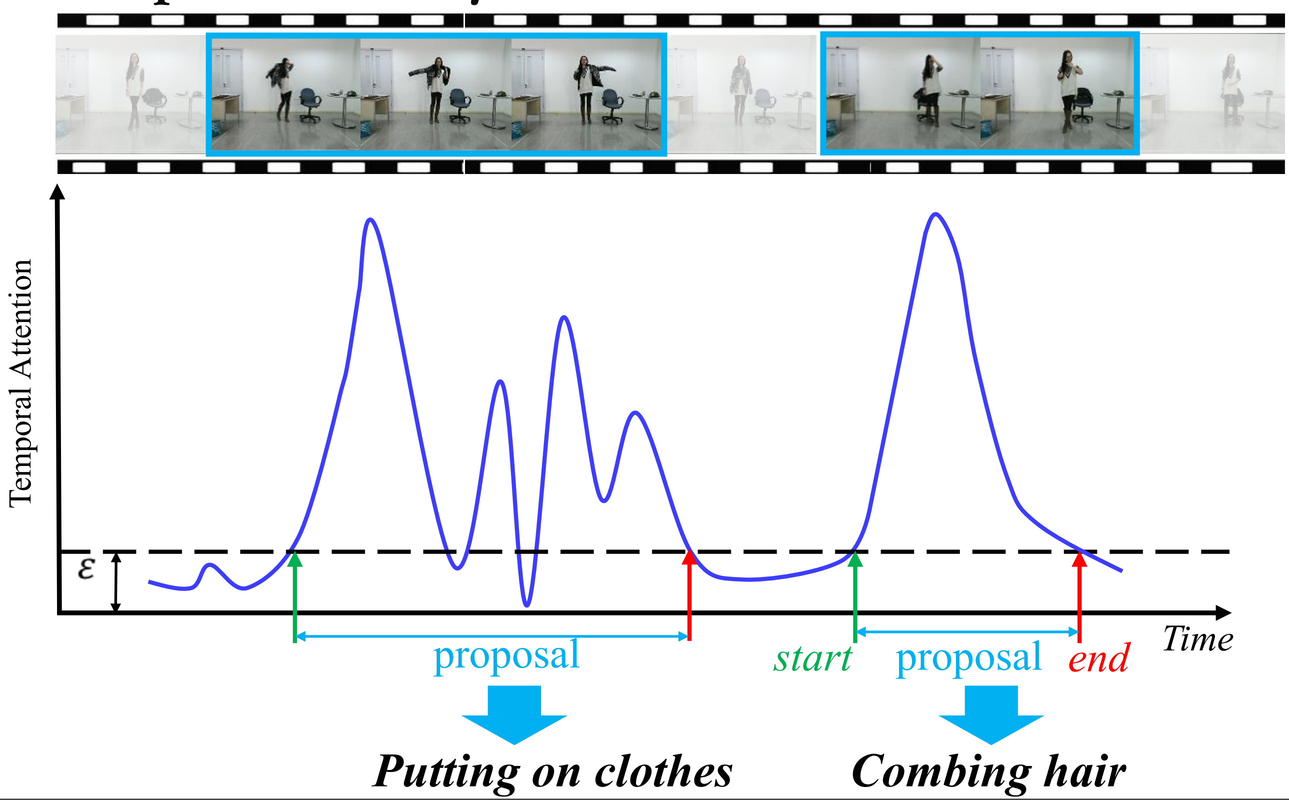
STA-LSTM as Classifier



Action Classification

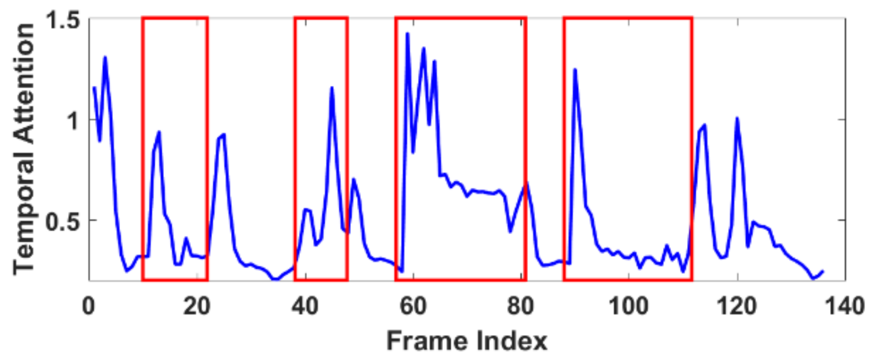
Classification on Each Proposal

- STA-LSTM as Classifier
- Post-processed by NMS

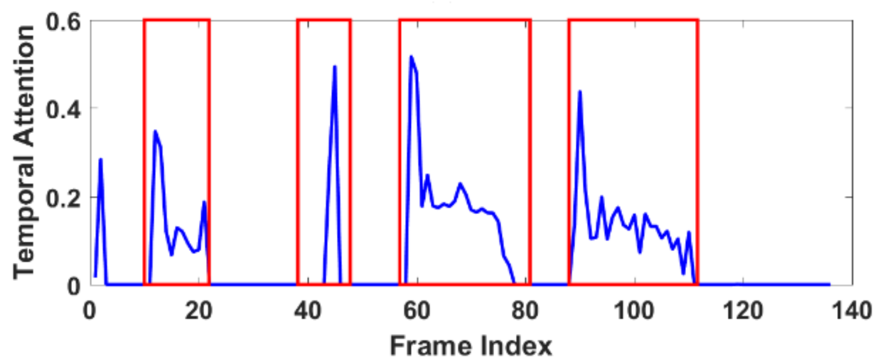


Experimental Results

- Visualization of Action Localization on PKUMMD dataset
  - LSTM vs. Bi-LSTM



(a) Temporal Attention from LSTM



(b) Temporal Attention from Bi-LSTM

Experimental Results

- Action detection results on PKUMMD dataset
  - TAP-U: Proposal generation by LSTM
  - TAP-U-M: Multiscale proposal generation by LSTM
  - TAP-B: Proposal generation by Bi-LSTM
  - TAP-B-M: Multiscale proposal generation by Bi-LSTM

Experimental Results

- Action detection results on PKUMMD dataset
  - mAP (mean Average Precision)

Method	Cross-Subject				Cross-View			
θ	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
SVM-SW	0.157	0.075	0.018	0.002	0.209	0.111	0.033	0.004
STA-LSTM-SW	0.416	0.298	0.129	0.015	0.425	0.306	0.148	0.027
JCR-RNN [Li 16]	0.499	0.440	0.345	0.169	0.612	0.525	0.417	0.222
STA-LSTM-JCR	0.504	0.465	0.364	0.177	0.584	0.540	0.425	0.176
TAP-U	0.398	0.348	0.241	0.074	0.419	0.361	0.228	0.052
TAP-U-M	0.466	0.410	0.216	0.047	0.522	0.470	0.292	0.079
TAP-B	0.483	0.461	**0.395**	**0.222**	0.572	0.530	0.450	0.255
TAP-B-M	**0.513**	**0.480**	0.352	0.155	**0.632**	**0.585**	**0.486**	**0.290**

Experimental Results

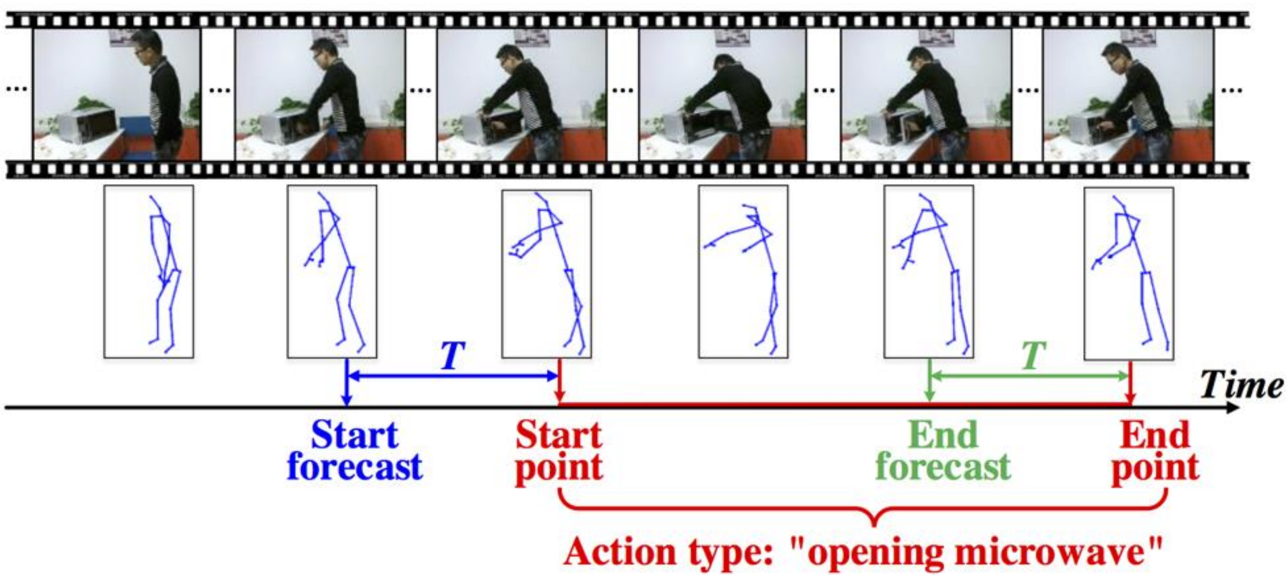
- Action detection results on PKUMMD dataset
  - F1-Score

Method	Cross-Subject				Cross-View			
θ	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
SVM-SW	0.266	0.168	0.074	0.010	0.322	0.220	0.108	0.026
STA-LSTM-SW	0.470	0.373	0.231	0.067	0.492	0.397	0.258	0.100
JCR-RNN [Li 16]	0.354	0.305	0.243	0.142	0.378	0.323	0.262	0.167
STA-LSTM-JCR	0.384	0.354	0.290	0.176	0.421	0.391	0.323	0.206
TAP-U	0.502	0.450	0.352	0.169	0.467	0.404	0.280	0.196
TAP-U-M	0.551	0.505	0.345	0.128	0.551	0.505	0.361	0.160
TAP-B	0.544	0.514	**0.461**	**0.327**	0.614	0.578	0.511	0.356
TAP-B-M	**0.557**	**0.530**	0.431	0.242	**0.651**	**0.608**	**0.531**	**0.385**

Online Action Detection & Forecasting

Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, Jiaying Liu. "Online Human Action Detection using Joint Classification-Regression Recurrent Neural Networks", *Proc. of by European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, Oct. 2016.

- **Action Detection**
  - Detect actions on the fly → Practical use
- **Action Forecasting**
  - Predict actions before they start/end → Better use experience



Skeleton-based Action Detection

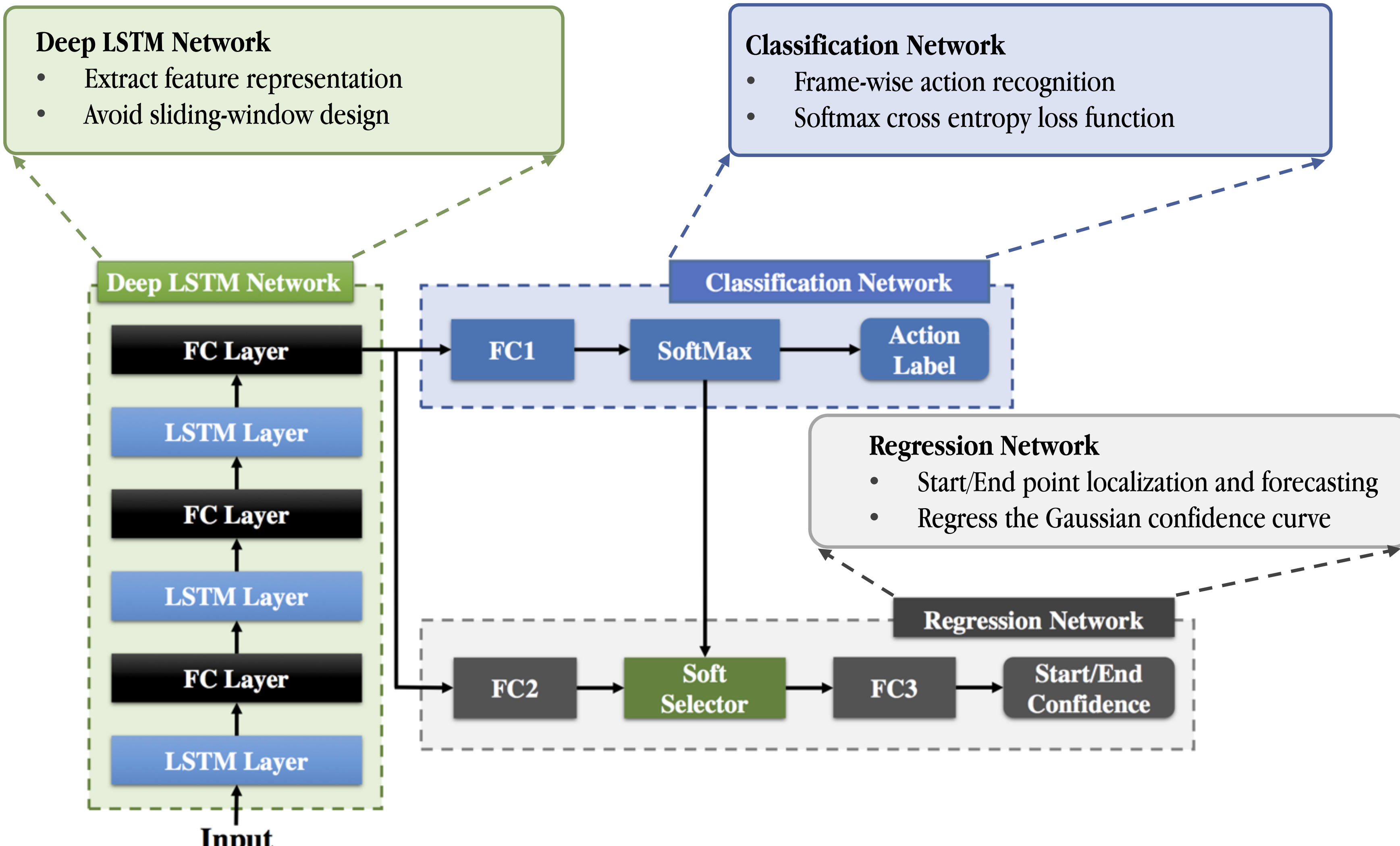
- Few skeleton-based detection works
- Offline Action Detection
  - Sliding window[Sharaf15]
  - Action proposal[Escorcia16]
  - → Time-consuming & Low precision
- Early Detection[Hoai14]
  - Recognize action after start

Skeleton-based Action Detection

- Few skeleton-based detection works
- Offline Action Detection
  - Sliding window[Sharaf15]
  - Action proposal[Escorcia16]
  - → Time-consuming & Low precision
- Early Detection[Hoai14]
  - Recognize action after start

- **Lack online action detection for practical use**

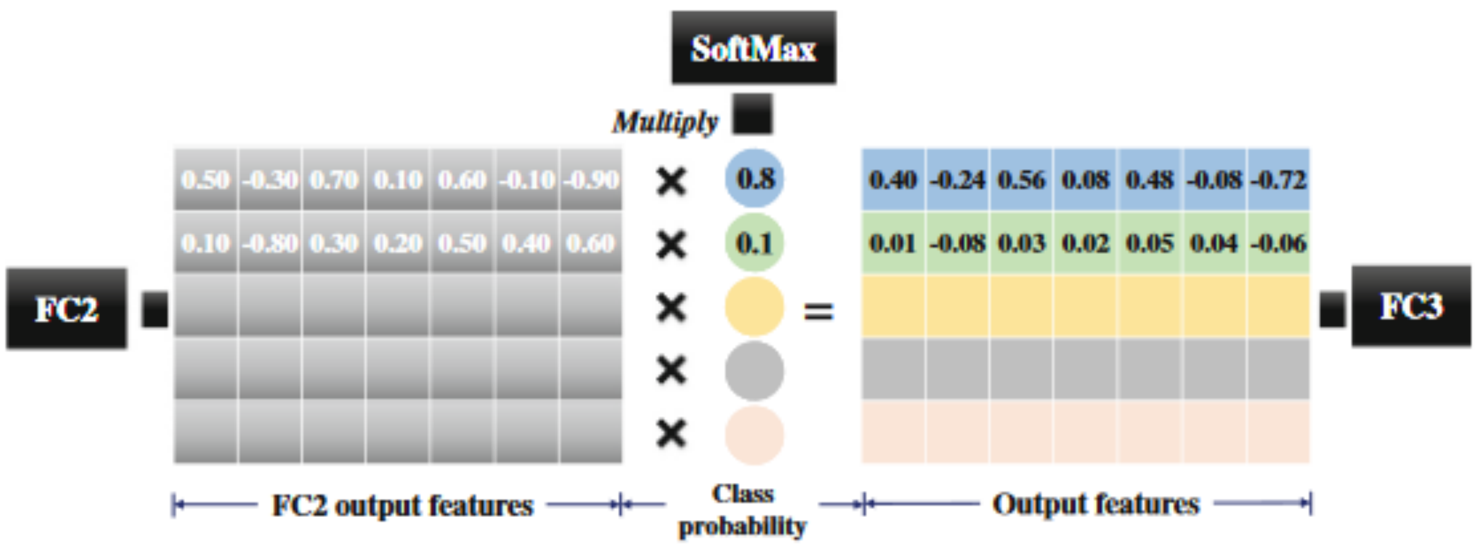
Joint Classification-Regression RNN



Joint Classification-Regression RNN

Soft-Selector Layer

- Incorporate classification info → Regression network
- Select class-specific features



Joint Classification-Regression RNN

Objective Function

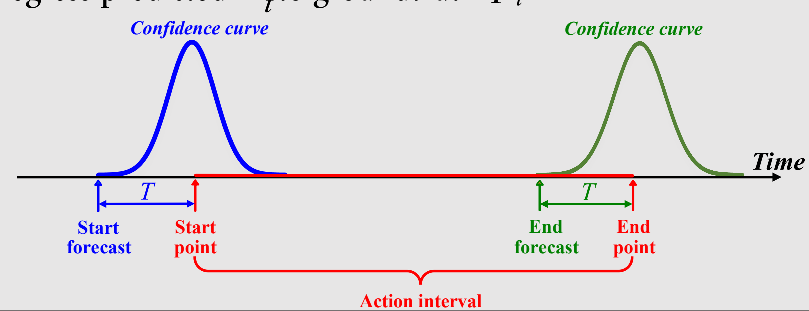
Classification Network Loss

- Cross entropy

$$\mathcal{L}(V) = -\frac{1}{N}\sum_{t=0}^{N-1}\left[\left(\sum_{k=0}^{M} z_{t,k}\ln P(y_{t,k}|v_0,...v_t)\right) + \lambda\cdot\left(\ell(c_t^s,p_t^s)+\ell(c_t^e,p_t^e)\right)\right]$$

Regression Network Loss

- Measure the possibility of being start/end point
- Regress predicted c_t^s to groundtruth p_t^s



Experimental Settings

- **Dataset**
  - Online Action Detection Dataset (OAD)
    - 59 sequences of 10 classes
    - 103347 frames
- **Evaluation criterions**
  - Action detection
    - F1-Score
    - SL/EL-Score
  - Action forecasting
    - Precision and Recall

Action Detection

Comparison with other methods

F1-Score

Actions	SVM -SW	RNN -SW	CA -RNN	JCR -RNN
drinking	0.146	0.441	**0.584**	0.574
eating	0.465	0.550	**0.558**	0.523
writing	0.645	**0.859**	0.749	0.822
opening cupboard	0.308	0.321	0.490	**0.495**
washing hands	0.562	0.668	0.672	**0.718**
opening microwave	0.607	0.665	0.468	**0.703**
sweeping	0.461	0.590	0.597	**0.643**
gargling	0.437	0.550	0.579	**0.623**
throwing trash	0.554	**0.674**	0.430	0.459
wiping	**0.857**	0.747	0.761	0.780
average	0.540	0.600	0.596	**0.653**

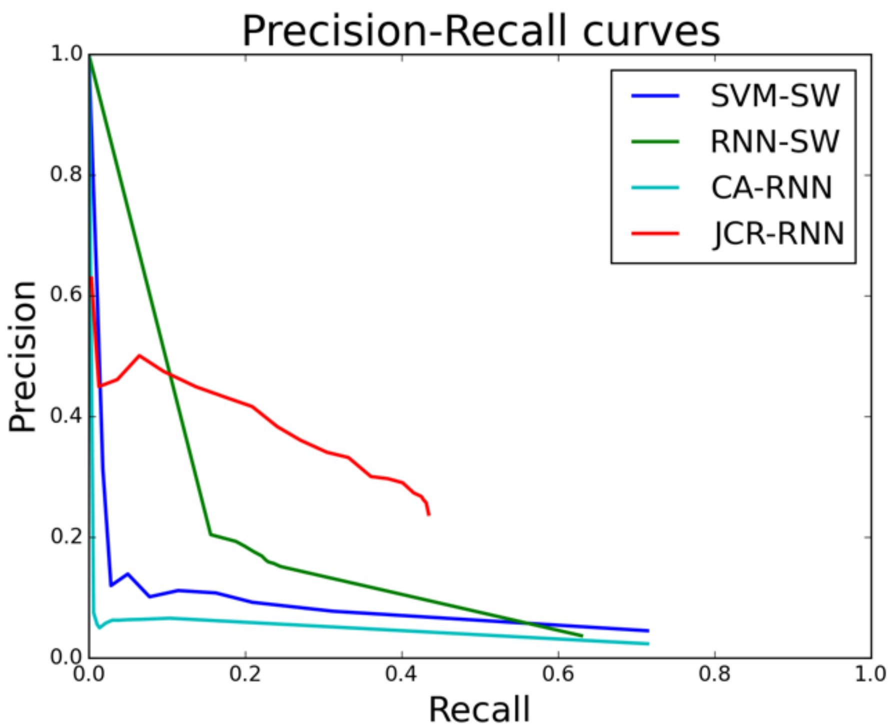
SL/EL-Score

Scores	SVM -SW	RNN -SW	CA -RNN	JCR -RNN
$SL-$	0.316	0.366	0.378	**0.418**
$EL-$	0.325	0.376	0.382	**0.443**

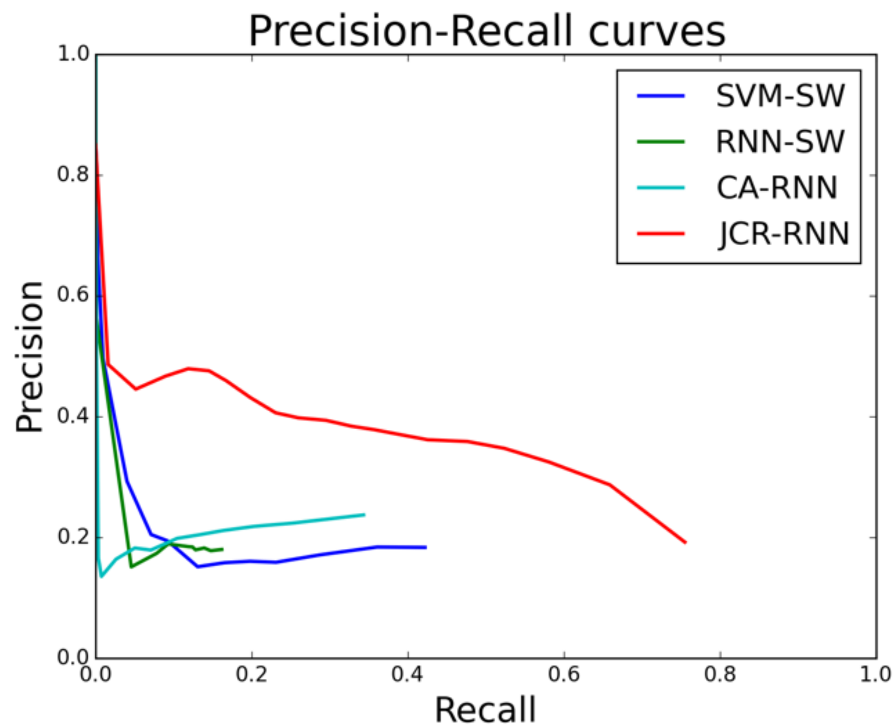
- SVM-SW
  SVM classifier with sliding window scheme
- RNN-SW
  RNN classifier with sliding window scheme
- CA-RNN
  Only with classification network
- JCR-RNN
  Proposed method

Action Forecasting

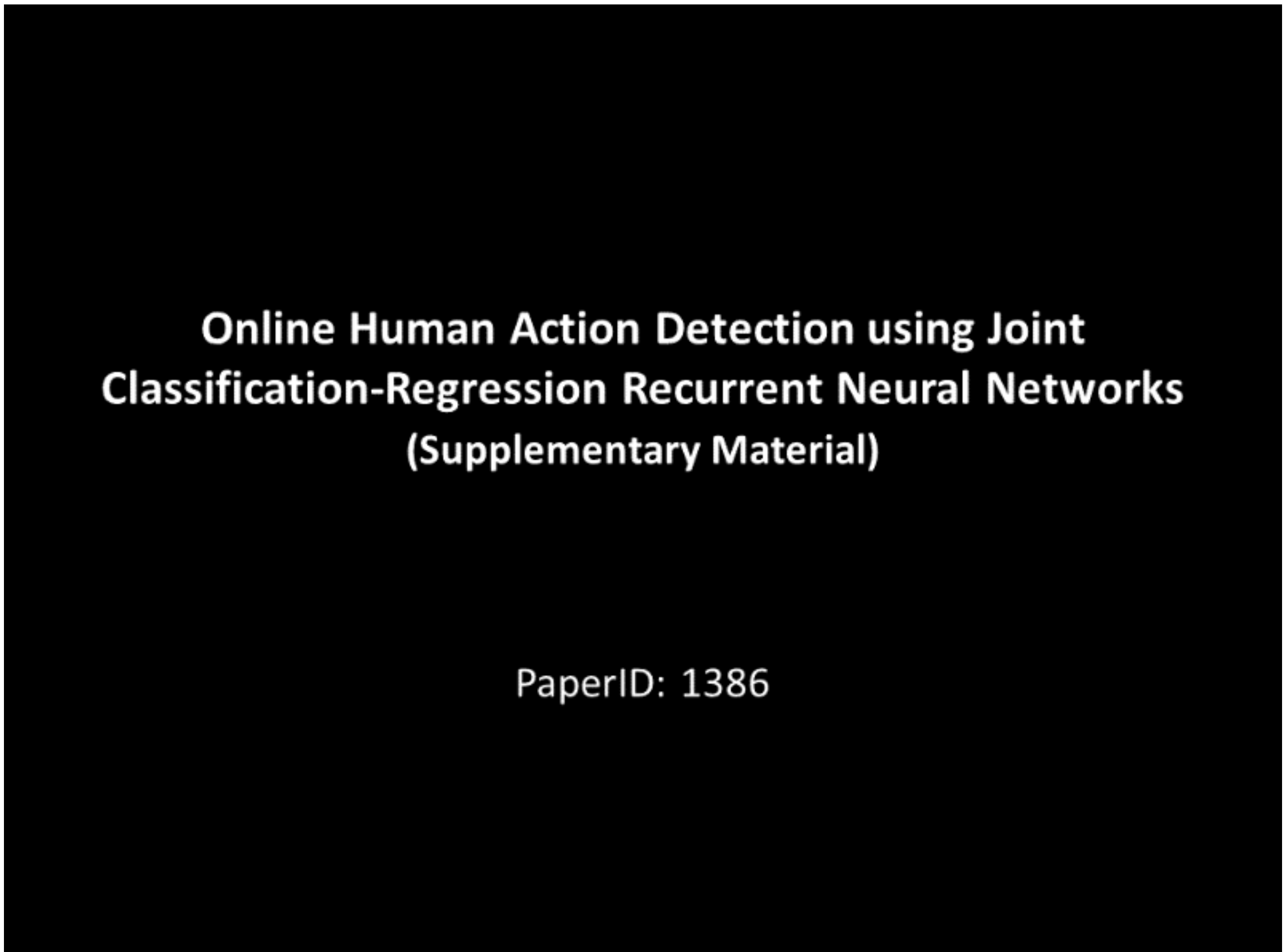
Forecast start point



Forecast end point

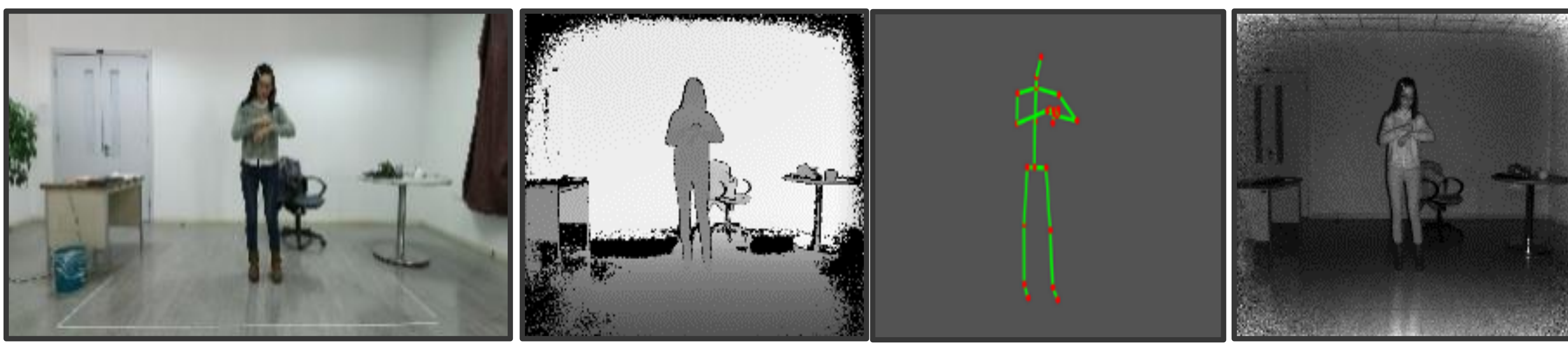


Demo


Online Human Action Detection using Joint
Classification-Regression Recurrent Neural Networks
(Supplementary Material)
PaperID: 1386

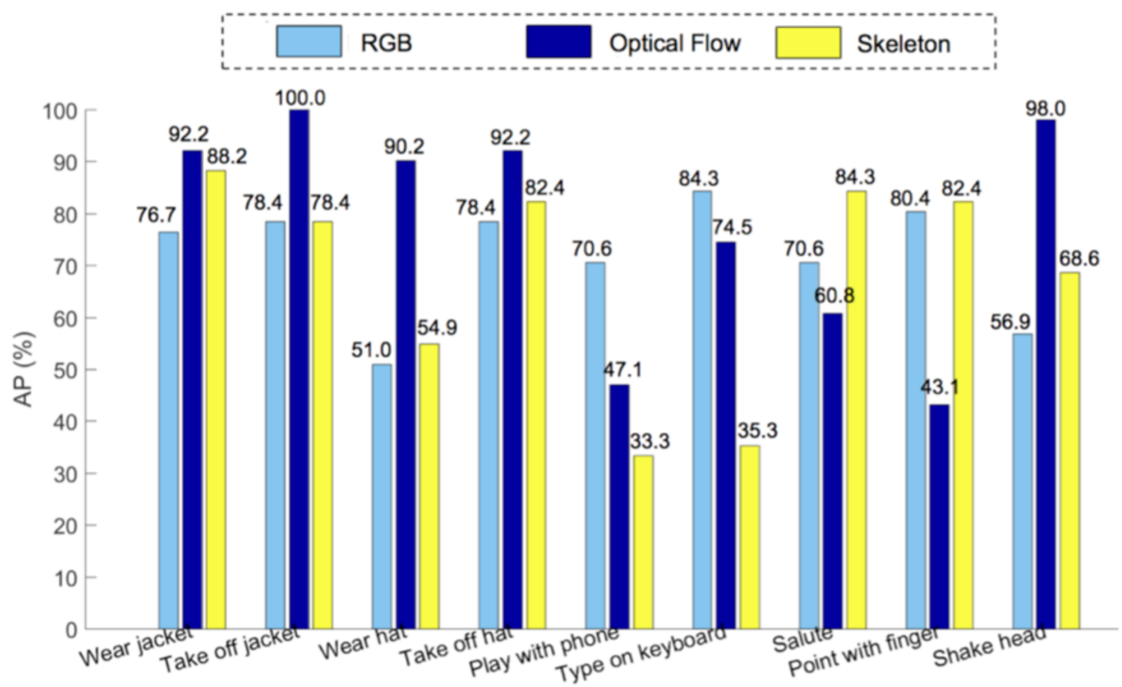
Multi-Modality Combination

- **Multi-Modality Action Detection and Forecasting**
  - Complementarity of different modalities
  - Characteristics of different modalities
  - → Multi-Modality Multi-task framework



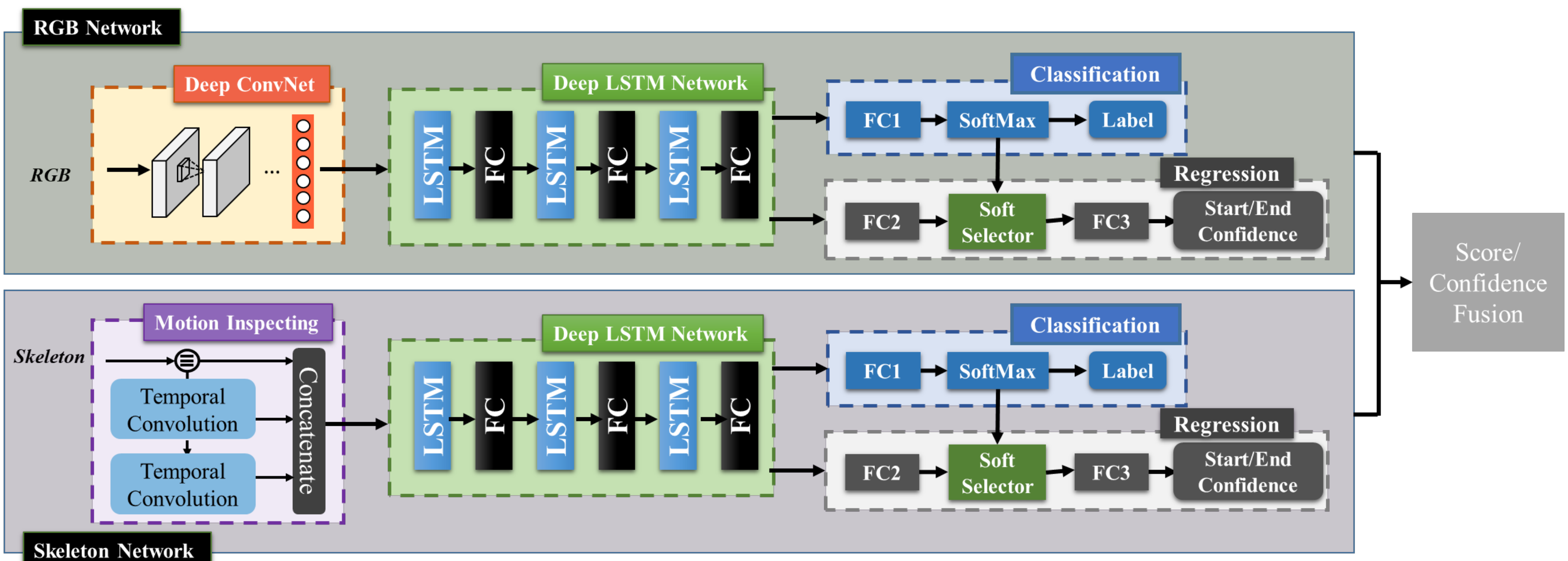
Multi-Modality Combination

- **Complementary of Modalities**
  - RGB → Rich appearance information
  - Optical flow → Motion information
  - Skeleton → high-level human representation



Multi-Modality Multi-Task RNN

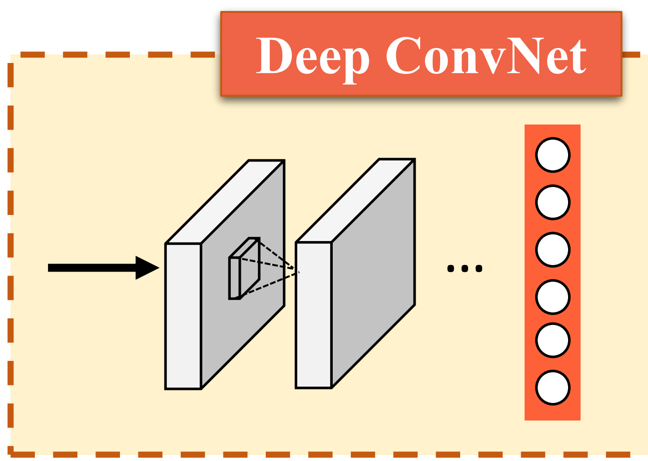
- Modality-specific temporal modeling network
- Deep LSTM network
- Joint Classification and Regression Model



Multi-Modality Multi-Task RNN

- Modality-specific temporal modeling network
  - RGB / Optical flow
    - Deep convolutional networks

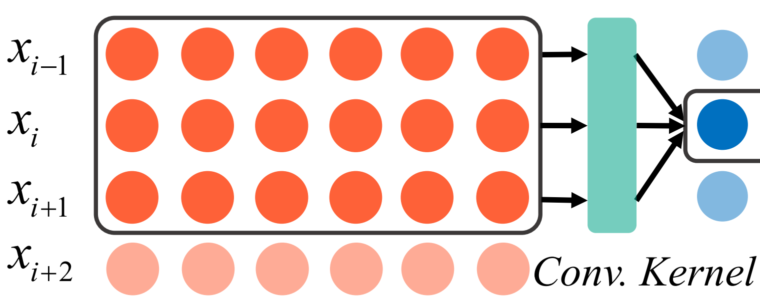
→ Extract high-level visual features

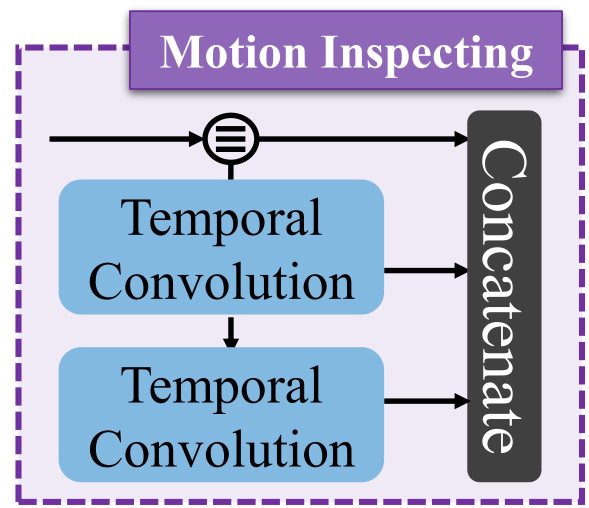


Multi-Modality Multi-Task RNN

- Modality-specific temporal modeling network
  - RGB / Optical flow
  - Skeleton data
    - Motion Inspecting layer

→ Extract high order motion features





Experimental Results

- Action detection results with different modalities
  - OAD dataset
  - Performance improved with multi-modality

Modality\Scores	$F1$-	SL-	EL-
RGB	0.578	0.341	0.371
Optical Flow	0.704	0.472	0.489
Skeleton	0.694	0.475	0.495
RGB + Optical Flow	0.741	0.491	0.517
RGB + Skeleton	0.692	0.442	0.470
Optical Flow + Skeleton	**0.795**	**0.576**	**0.597**
RGB + Optical Flow + Skeleton	0.785	0.561	0.577

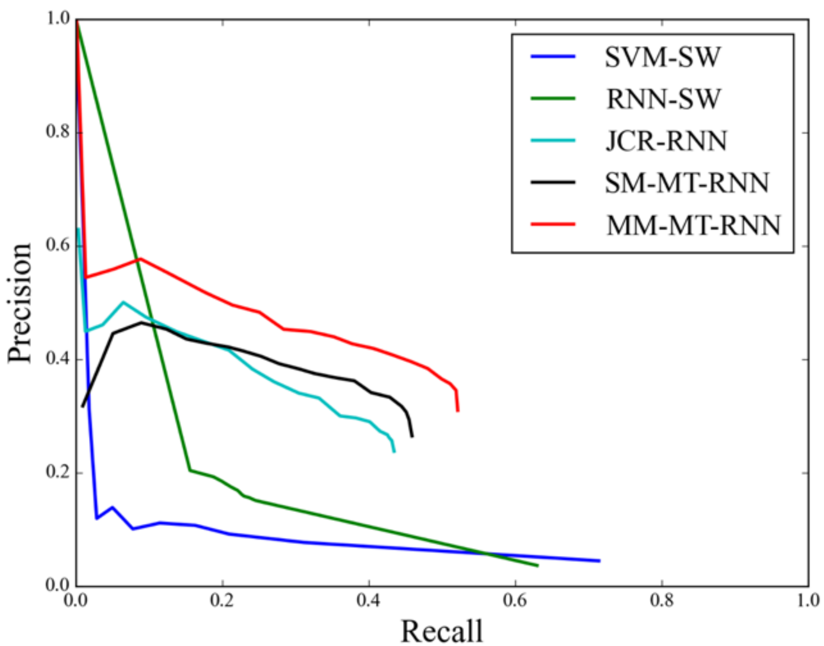
Experimental Results

- Action detection results with different methods
  - OAD dataset

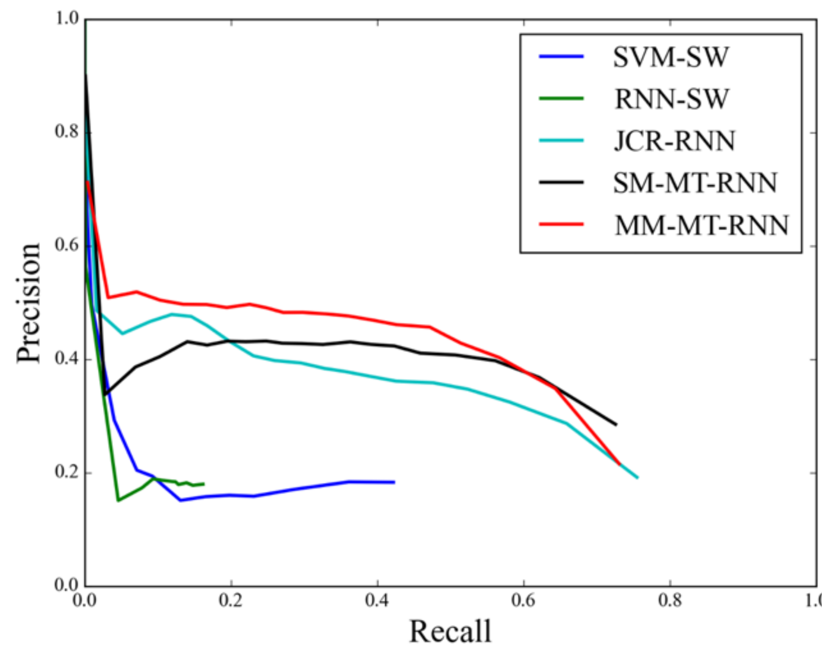
Modalities	Method \Scores	$F1$-	SL-	EL-
Skeleton	SVM-SW	0.540	0.316	0.325
	RNN-SW[Zhu16]	0.600	0.366	0.376
	JCR-RNN[Li16]	0.653	0.418	0.443
	SM-MT RNN	**0.694**	**0.475**	**0.495**
Multi-modality	RF+ST[Baek17]	0.672	0.445	0.432
	MM-ST RNN	0.748	0.515	0.522
	MM-MT RNN	**0.795**	**0.576**	**0.597**

Action Forecasting

- **Forecast start point**



- **Forecast end point**



Intelligent Action Analytics

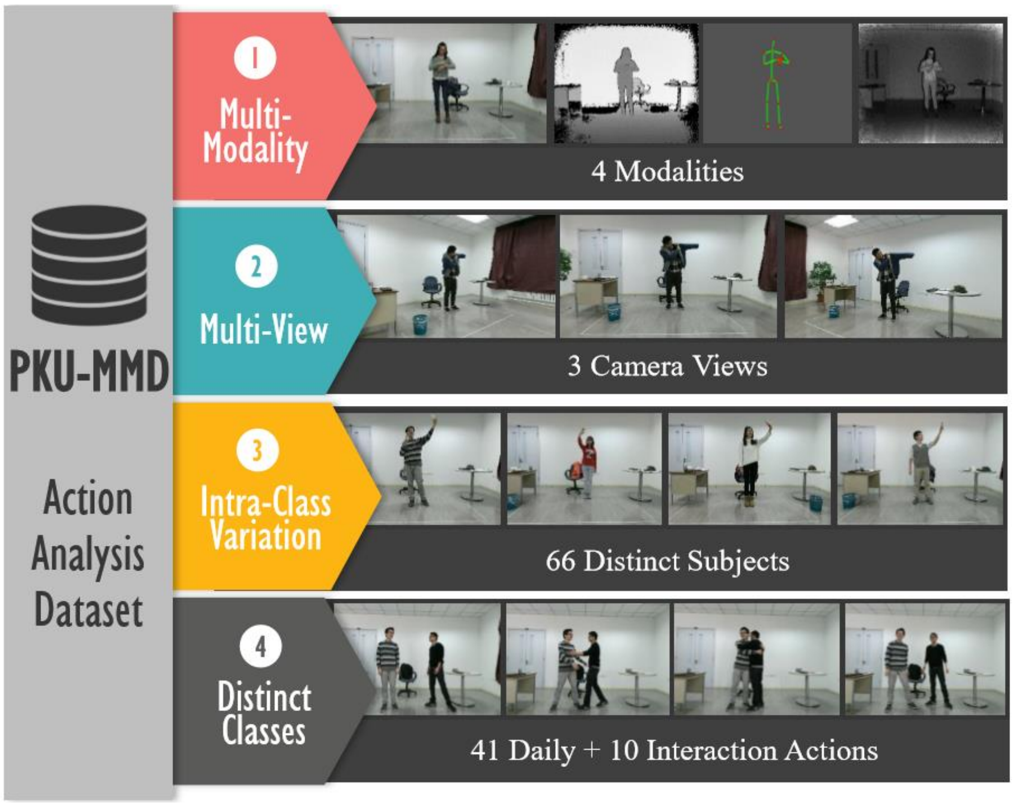
Outline

Data Sources

Compared with Other Datasets

Datasets	Classes	Videos	Labeled Instances	Actions per Video	Modalities	Year
G3D[Bloom 2012]	20	210	1467	7	RGB+D+Skeleton	2012
SBU [Yun, 2012]	8	21	300	14.3	RGB+D+Skeleton	2012
CAD-120[Sung, 2012]	20	120	~1200	~8.2	RGB+D+Skeleton	2013
compostable Activities[Lilo, 2014]	16	693	2529	3.6	RGB+D+Skeleton	2014
Watch-n-Patch[Wu, 2015]	21	458	~2500	2~7	RGB+D+Skeleton	2015
OAD[Li, 2016]	10	59	~700	~12	RGB+D+Skeleton	2016
PKU-MMD	**51**	**1076**	**21545**	**20.02**	**RGB+D+Skeleton +IR**	**2017**

PKU-MMD



- **Large scale**
  - Over 5 million frames
  - Over 3000 minutes
- **Variation**
  - Multi-modality
  - 3 camera views
  - 66 subjects
  - 51 action categories

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. "PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding", Proc. of *ACM Multimedia* (*ACM MM*) *Workshop*, Silicon Valley, California, U.S., Oct. 2017

PKU-MMD

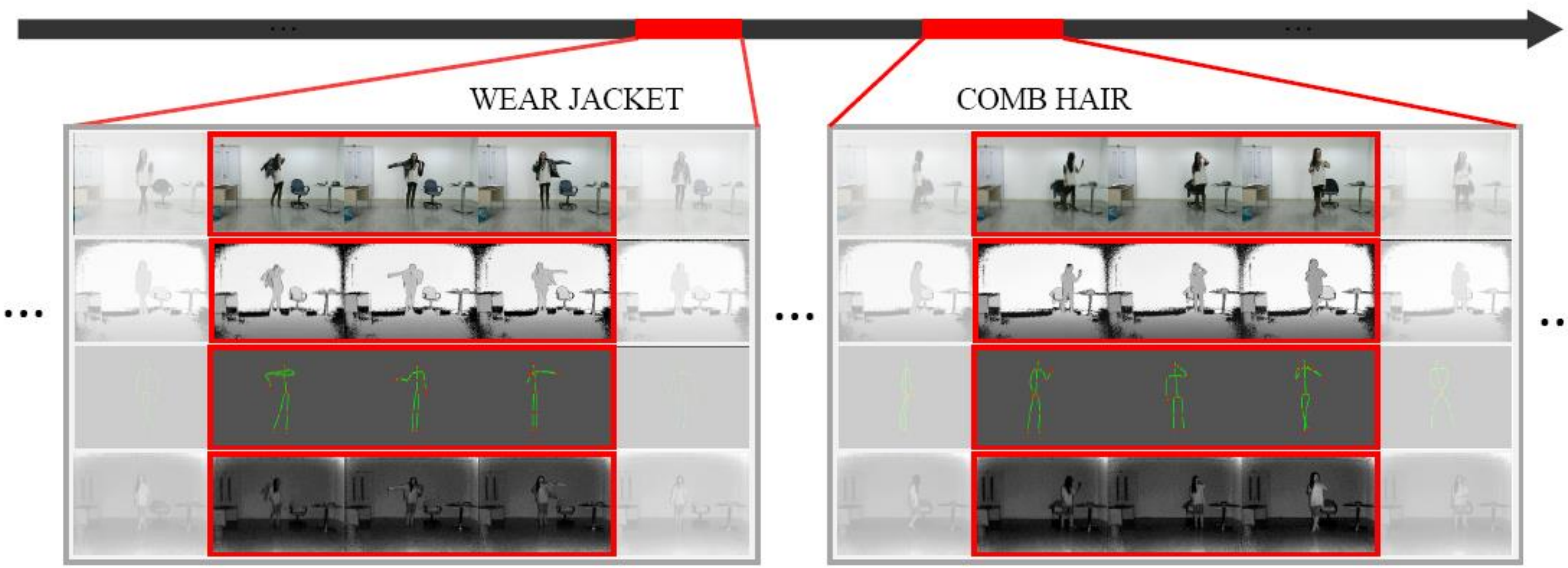
- **Action Types**
  - Health related: touch head (headache), touch chest (heart attack), …
  - Home related: brush teeth, comb hair, …
  - Dressing related: put on a jacket, take off glasses, …
  - Interaction with people: handshake, hug, …
  - Interaction with items: write, take a selfie, …
  - Human locomotion: clap, bow, …

PKU-MMD

- Action Recognition
  - 21545 labeled instances

PKU-MMD

- Action Detection
  - 1076 video samples

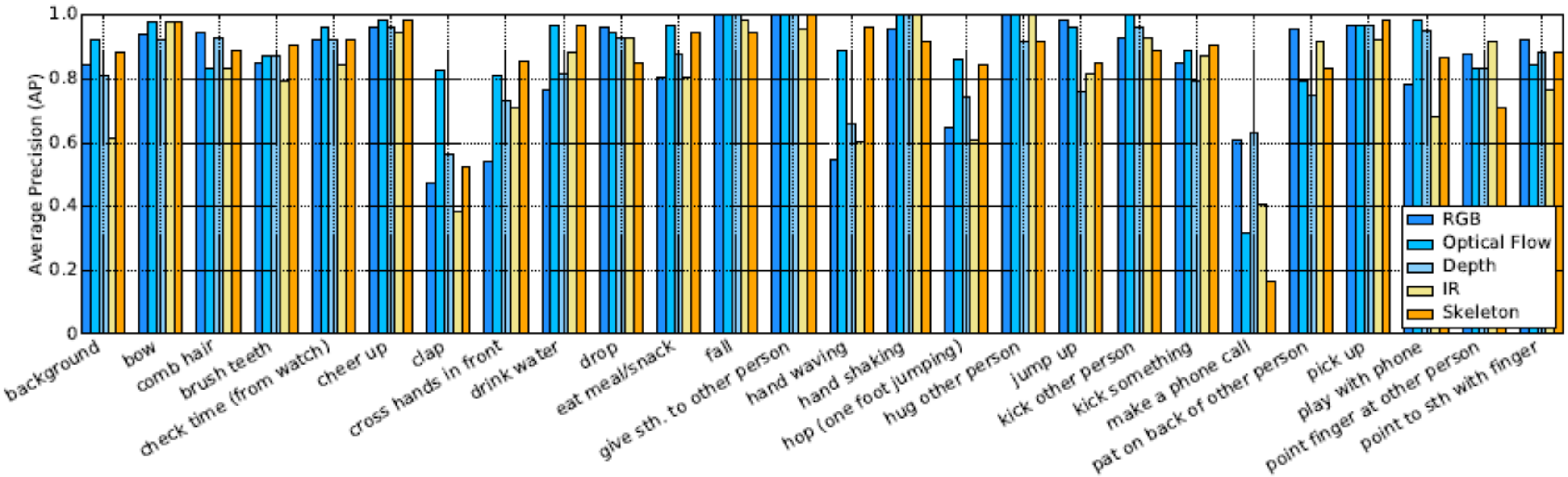


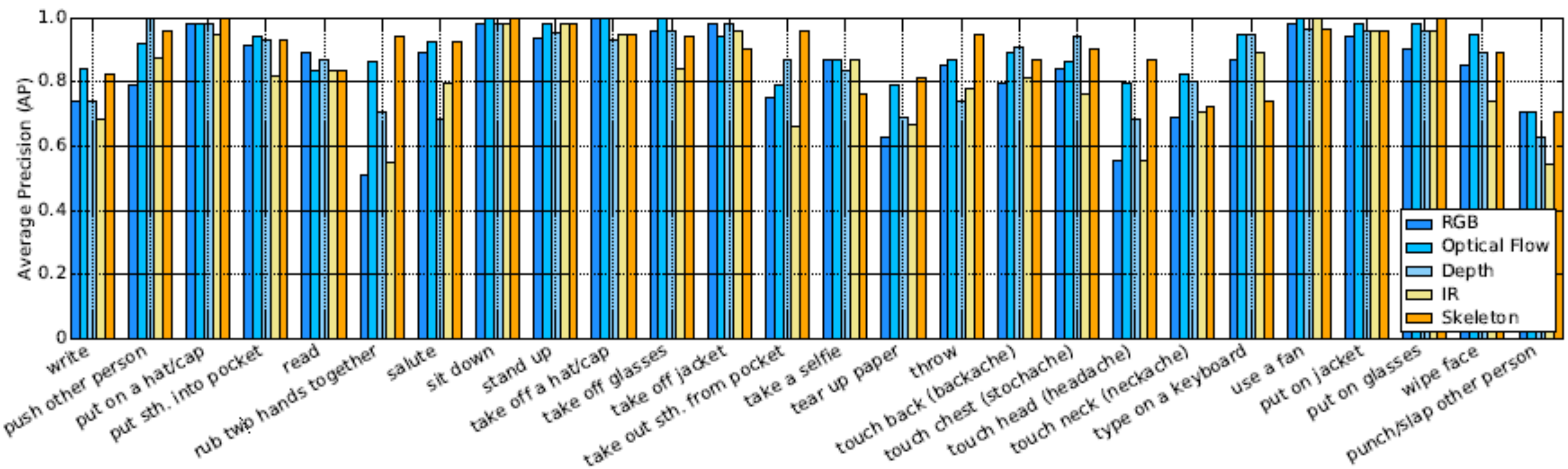
Benchmarks for Action Recognition

Results with different modalities

Methods (Acc. %)	Cross-Subject					Cross-View				
	R	F	D	IR	S	R	F	D	IR	S
TSN	79.03	87.86	79.00	73.16	--	84.13	90.26	76.85	67.95	--
STA-LSTM [Song 2017]	--	--	--	--	87.29	--	--	--	--	90.83
TPN[Hu, 2017]	--	--	--	--	85.71	--	--	--	--	93.71
LSTM	84.01	90.00	85.11	80.42	86.67	88.56	92.12	81.16	80.10	93.97
BLSTM	84.77	90.10	86.16	80.61	86.41	88.63	93.14	83.68	80.12	94.58

Benchmarks for Action Recognition





Benchmarks for Action Recognition

- Fusion results with different modalities (BLSTM)

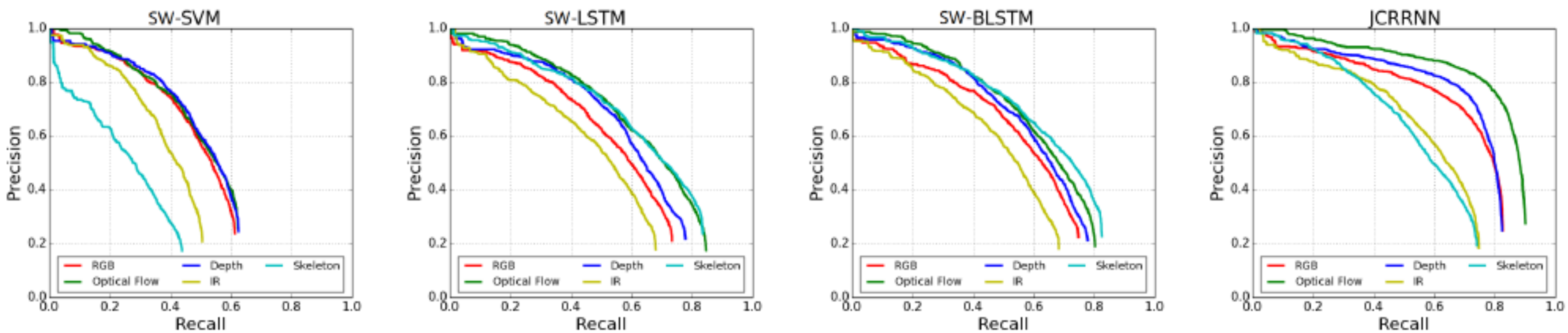
Modalities (Acc. %)	Cross-Subject	Cross-View
R+F	91.54	95.08
R+D	88.35	90.83
R+IR	85.20	89.20
R+S	90.22	96.18
R+F+D	92.27	95.04
R+F+IR	90.92	94.70
R+F+S	93.30	97.26
R+F+D+S	**94.36**	**97.45**
R+F+IR+S	93.00	96.89
R+F+D+IR+S	93.66	96.84

Benchmarks for Action Detection

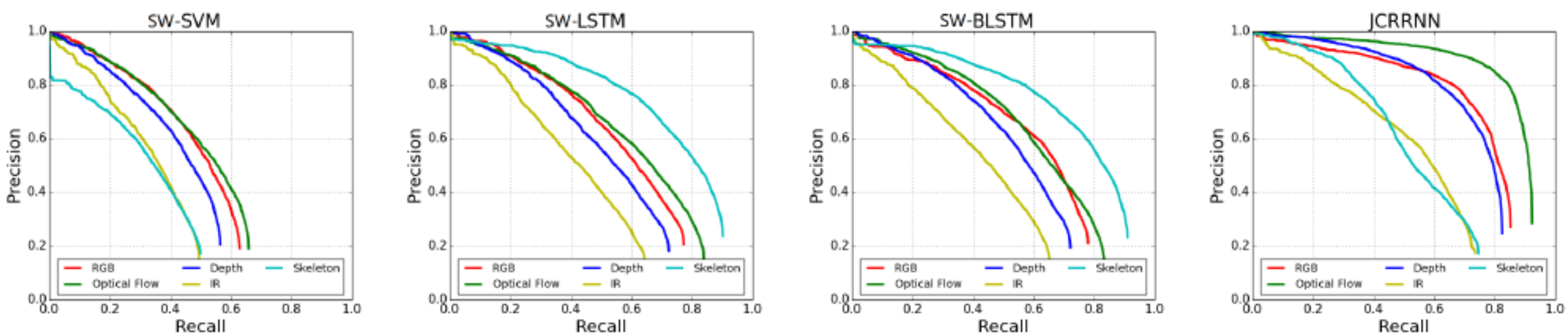
- Results with different modalities

Methods (mAP)	Cross-Subject					Cross-View				
	R	F	D	IR	S	R	F	D	IR	S
STA-LSTM-SW	--	--	--	--	0.254	--	--	--	--	0.278
TPN-SW	--	--	--	--	0.304	--	--	--	--	0.400
SVM-SW	0.236	0.188	0.252	0.173	0.131	0.200	0.146	0.179	0.131	0.152
LSTM-SW	0.304	0.254	0.348	0.256	0.382	0.305	0.201	0.263	0.195	0.449
BLSTM-SW	0.333	0.244	0.358	0.252	0.363	0.314	0.228	0.266	0.196	0.442
JCRRNN[Li, 2016]	0.538	0.668	0.579	0.428	0.355	0.615	0.742	0.576	0.402	0.380

Benchmarks for Action Detection



(a) Cross-subject



(b) Cross-view

Benchmarks for Action Detection

- Fusion results with different modalities (SW-BLSTM)

Modalities (mAP)	Cross-Subject		Cross-View	
θ	0.1	0.5	0.1	0.5
R+F	0.737	0.380	0.757	0.389
R+D	0.699	0.420	0.685	0.350
R+IR	0.654	0.356	0.652	0.316
R+S	0.755	0.450	0.815	0.486
R+F+D	0.767	0.458	0.764	0.422
R+F+IR	0.736	0.408	0.739	0.400
R+F+S	**0.808**	0.483	**0.849**	**0.541**
R+F+D+S	0.806	**0.516**	0.836	0.531
R+F+IR+S	0.791	0.481	0.825	0.514
R+F+D+IR+S	0.796	0.496	0.810	0.498