北京郵電大學

Beijing University of Posts and Telecommunications

BUPT-MCPRL@TRECVID 2019

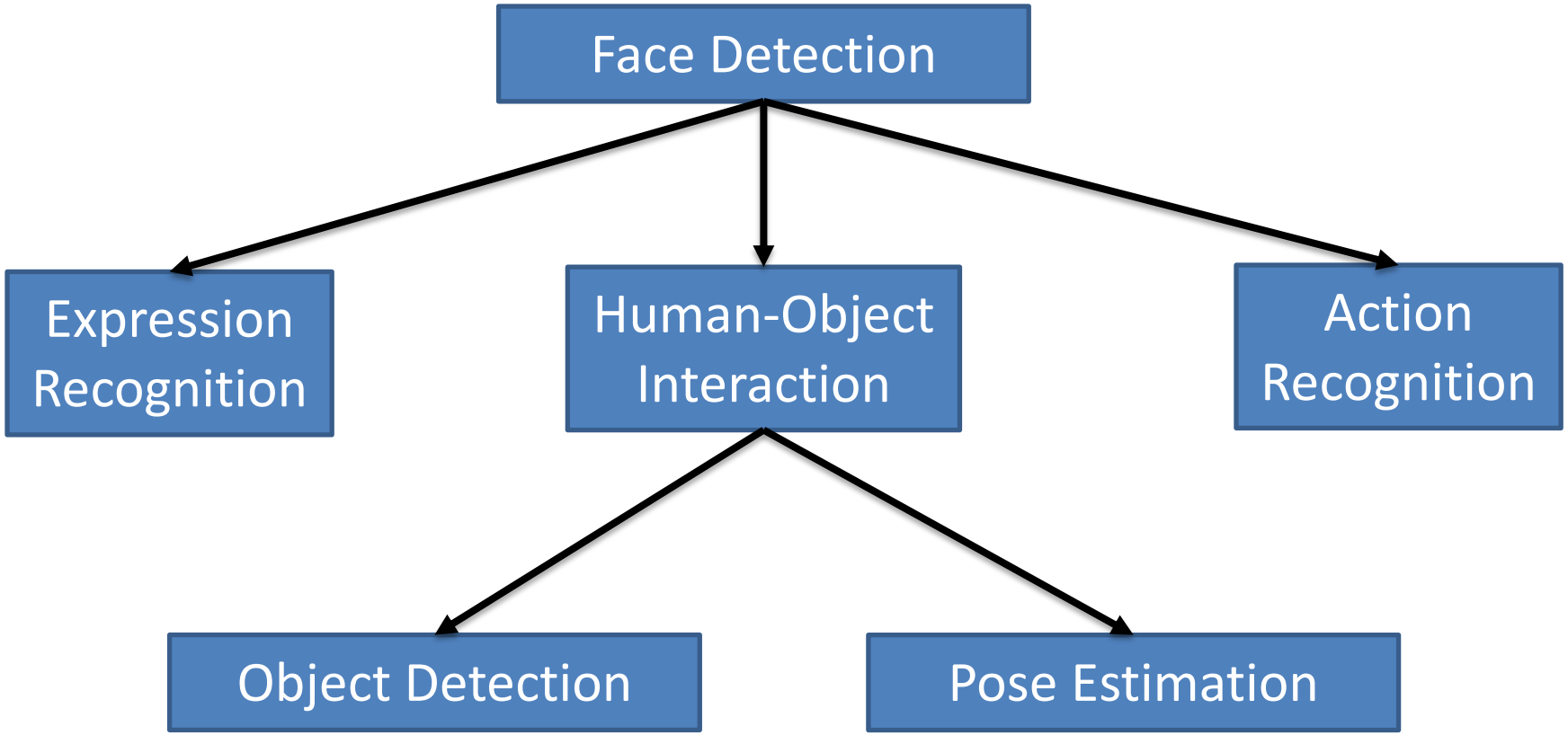
Guanyu Chen
Chong Chen, Xinyu Li, Xuanli Xiang
Zhicheng Zhao, Yanyun Zhao, Fei Su
Multimedia Communication and Pattern Recognition Labs,
Beijing University of Posts and Telecommunications
(BUPT-MCPRL)

loraschen@bupt.edu.cn

Instance Search

- Parse INS into multiple related visual subtasks, and propose a novel INS framework based on multi-task retrieval and re-ranking.
- An improved two-pathway ECO network (IECO) is designed to enhance video feature extraction.
- A new relative pose representation (RPR) is presented, and a light pose-based action recognition network is constructed to restrain the impacts of camera movement.
- The experimental results on four datasets demonstrate the effectiveness of the proposed INS framework.

Instance Search



Face Detection

Face Detection

crop

Face Features Extractor

Face Feature

dot product

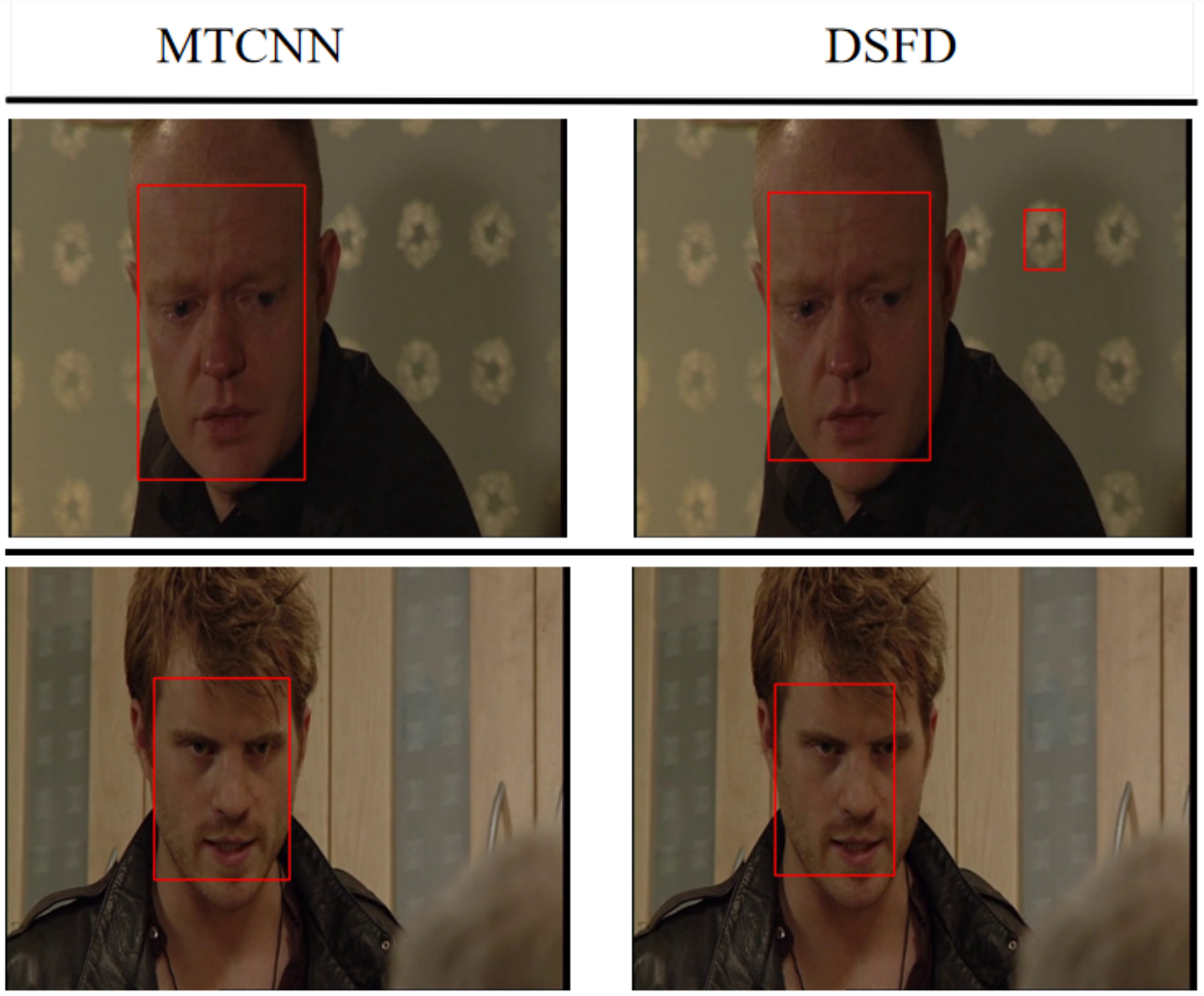
Cosine Similarity

Face Feature

Query Image

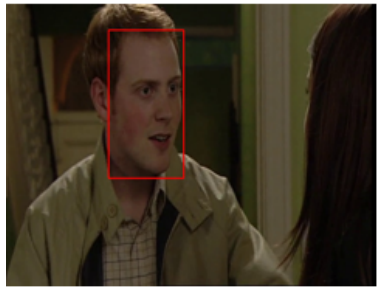
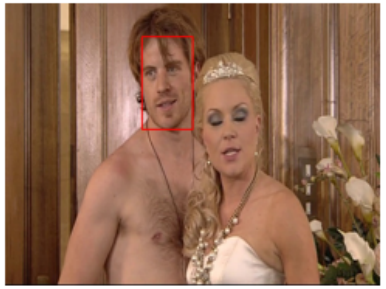
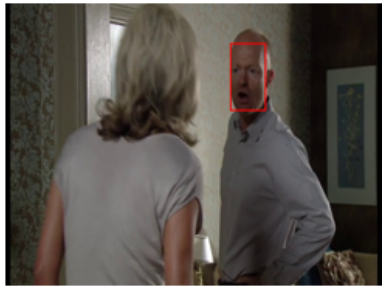
detect and crop

Face Detection



Compare MTCNN with DSFD. DSFD model could detect wrong faces and the detected bounding boxes is not exactly accurate sometimes .

Face Detection

1st	1000th	3000th	5000th	10000th
				
				
				

Expression Recognition

Upper image: Architecture of expression-related action retrieval.

MODEL_STRATEGY	FER2013 Testsets（Accuracy）
VGG19_SOFTMAX	68.89%
VGG19_DROPOUT_RANDOMCROP_SOFTMAX	71.49%

Lower table: Accuracy on public dataset FER2013.

Expression Recognition

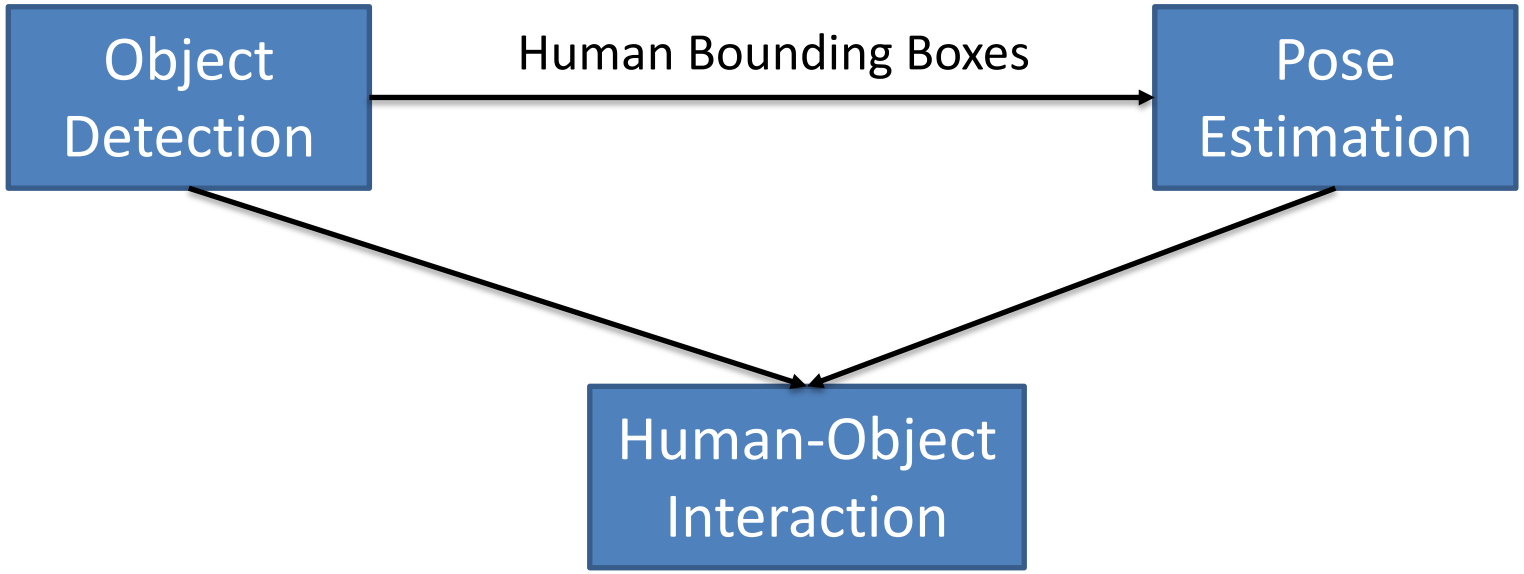
False Detection

Laughing

Crying

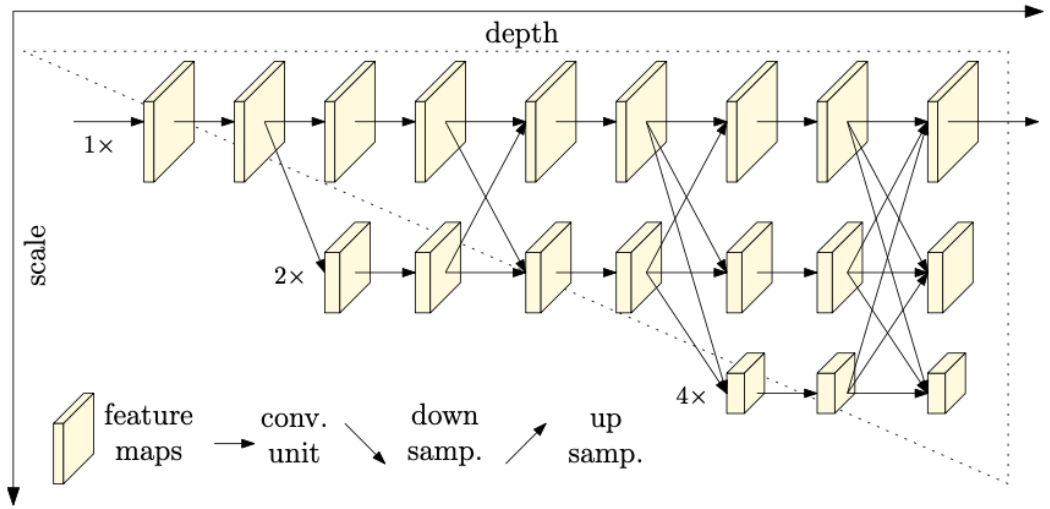
Shouting

Human-Object Interaction



1) Using YOLOv3 to detect key objects such as glass, bag, phone, person.
2) Feed human bounding boxes into HRNet to estimate human poses.
3) Calculate the relative distance between key objects and interactive keypoint to measure the dependences of human-object interaction and group the initial ranklist.

Human-Object Interaction



Left: Architecture of HRNet[1]. It could extract high-resolution representation from input image.

Right: Comparison of OpenPose and HRNet. The former method performs poorly when one person overlaps with another.

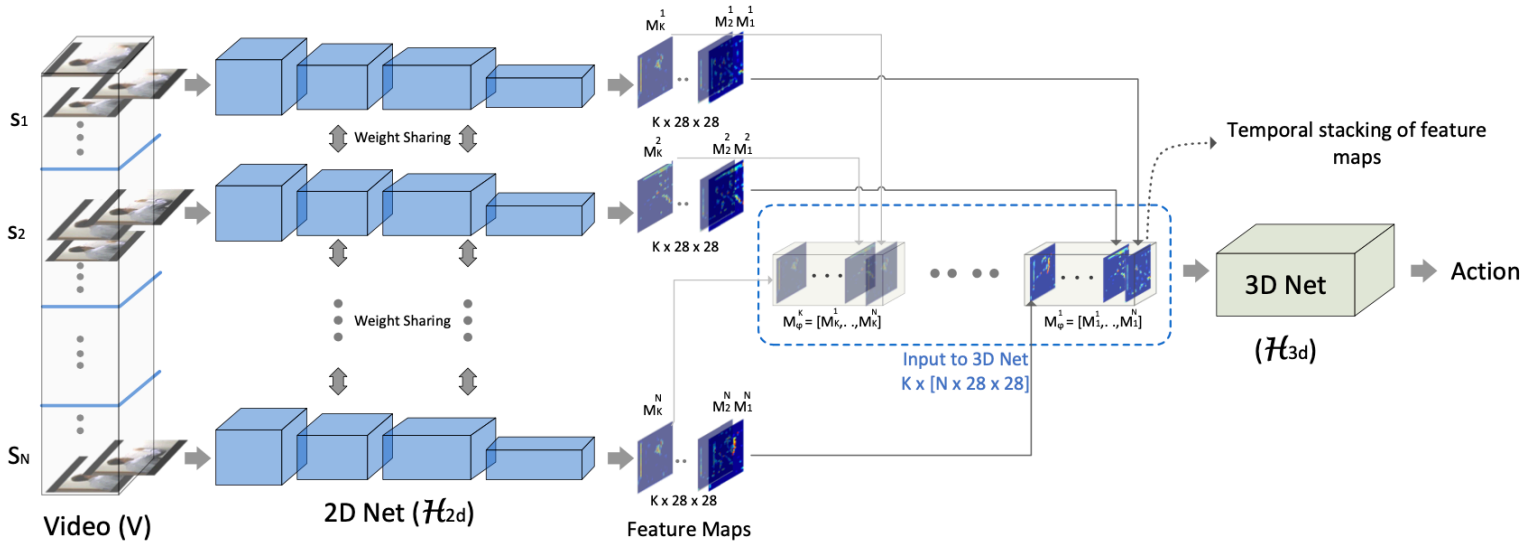
[1] Sun, Ke, et al. “Deep High-Resolution Representation Learning for Human Pose Estimation.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

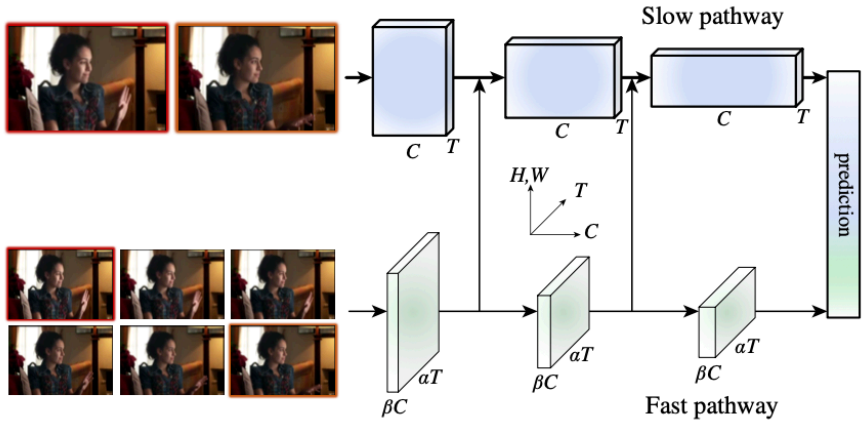
Human-Object Interaction

Pat Sit_on_couch

Ian Holding_phone

Action Recognition



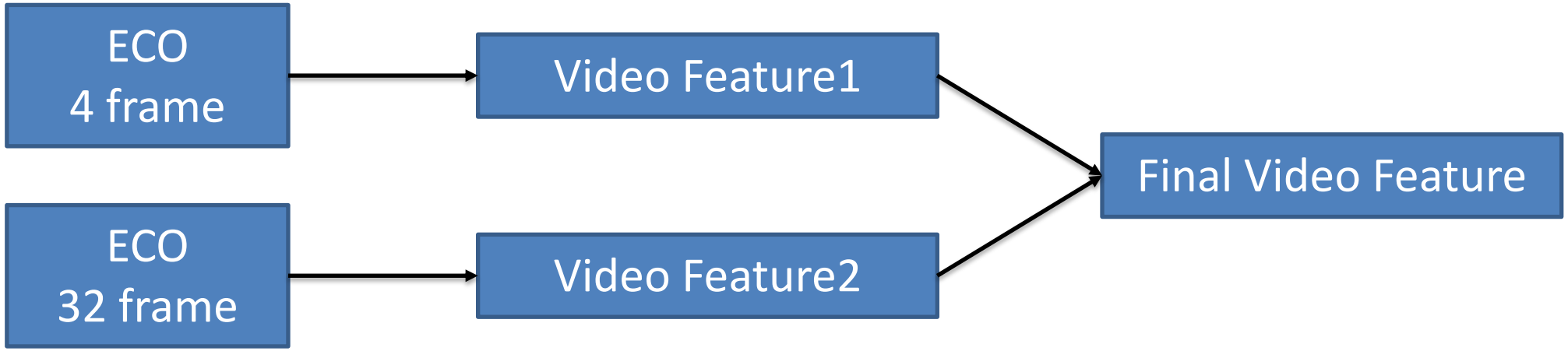


Left: Architecture of ECO[1], we choose it as the basic network for video vector extraction.

Right: Architecture of SlowFast[2], taking videos with different frame rates as input.

[1] Zolfaghari, Mohammadreza, Kamaljeet Singh, and Thomas Brox. "Eco: Efficient convolutional network for online video understanding." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
[2] Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." *arXiv preprint arXiv:1812.03982* (2018).

Action Recognition



Upper framework: Architecture of proposed IECO.

Pathway	HMDB(mAP)	UCF101(mAP)
One(16 frame)	46.68	67.90
Two(4 & 32 frame)	**54.39**	**72.89**

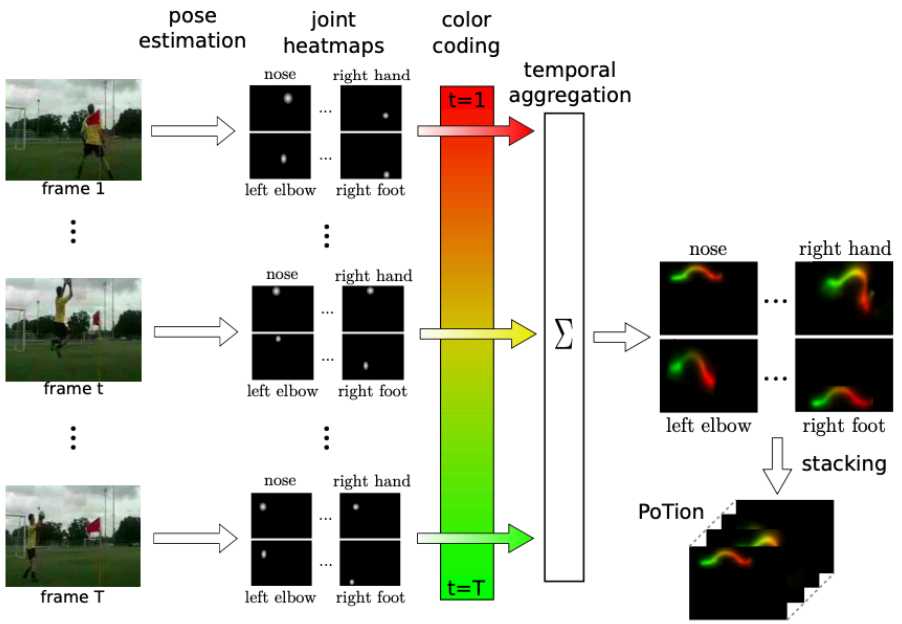
Lower table: Results on HMDB and UCF101 based on ECO with different pathways. It shows improvement of IECO on both two datasets.

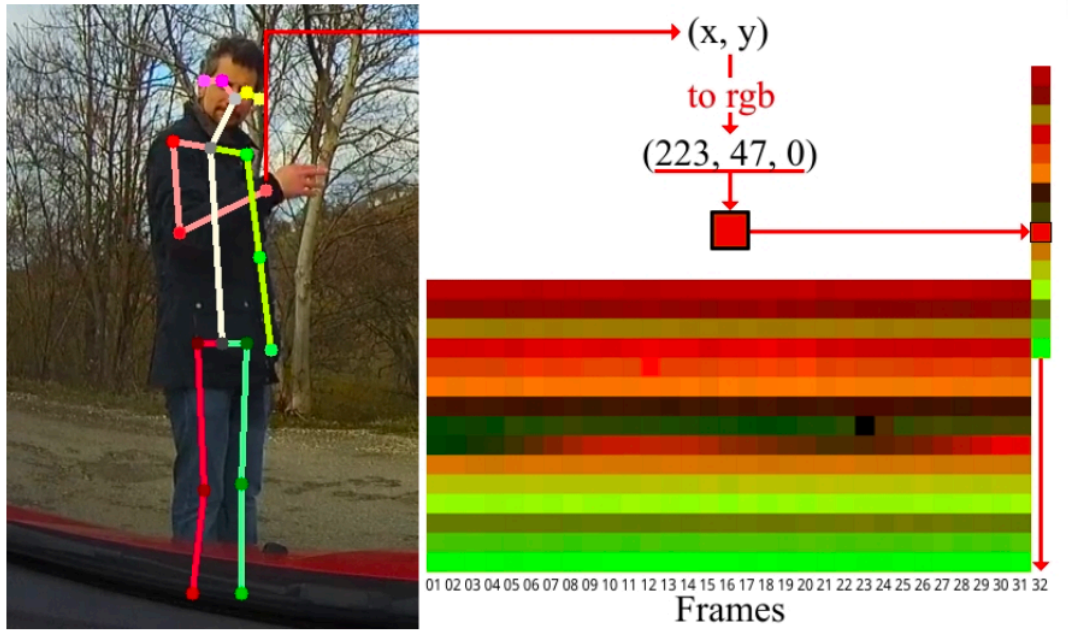
Action Recognition

Jack Kissing

Stacey Hugging

Pose-based Action Detection



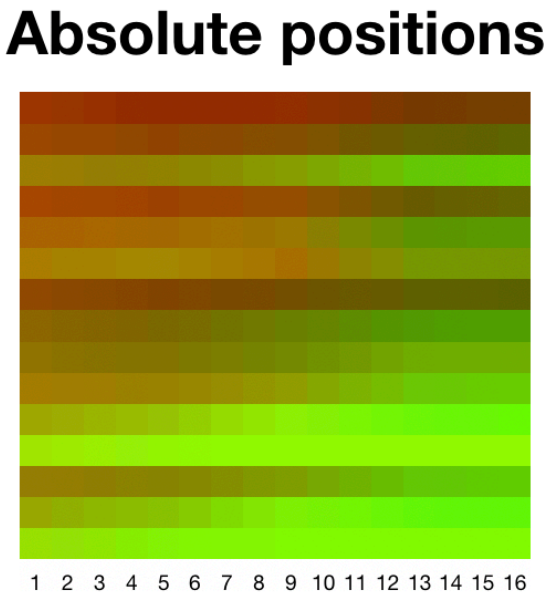


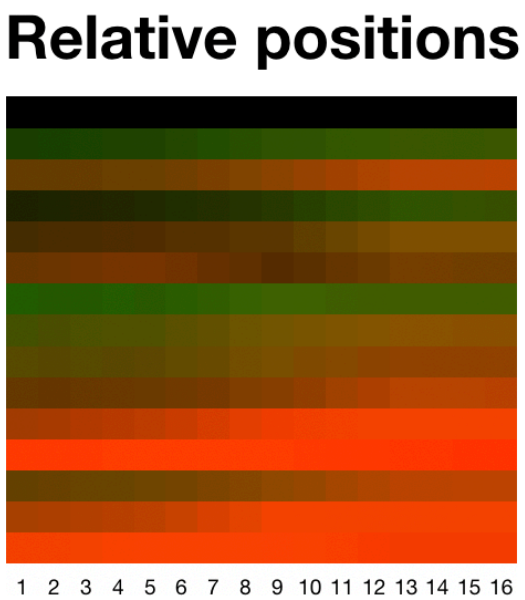
Two types of pose-based action detection models. The left[1] encodes the time information of keypoints motion, and the right[2] encodes the position of keypoints in the image.

[1] Choutas, Vasileios, et al. “Potion: Pose motion representation for action recognition.” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[2] Ludl, Dennis, Thomas Gulde, and Cristóbal Curio. "Simple yet efficient real-time pose-based action recognition." *arXiv preprint arXiv:1904.09140* (2019).

Pose-based Action Detection

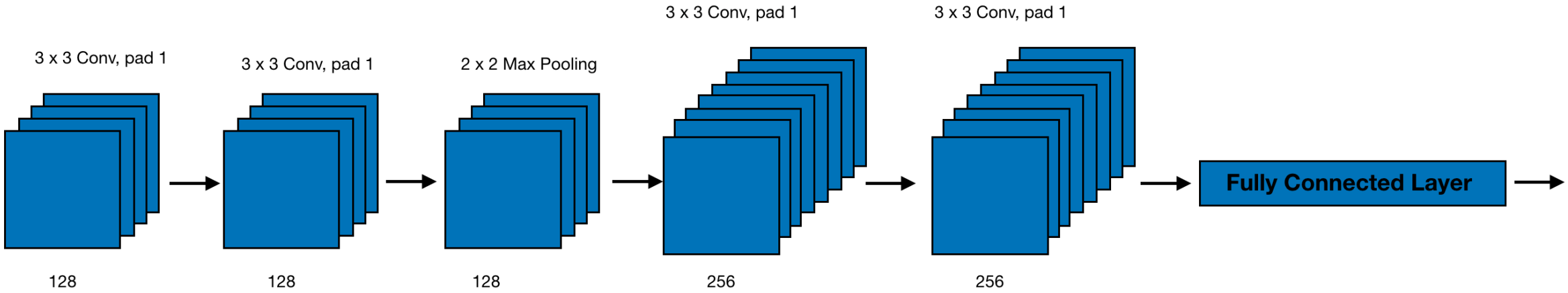




$rela\_dis$: normalized distance between keypoint and nose.
$rela\_angle$: the angle between x-axis and the line that joins keypoint and nose.

$$rela\_dis_i = \frac{dis_i}{\max\left(dis_1, dis_2, ..., dis_k\right)} \quad rela\_angle_i = \begin{cases} \frac{\arccos \frac{x_i - x_n}{dis_i}}{360} & y_i \geq y_n \\ 1 - \frac{\arccos \frac{x_i - x_n}{dis_i}}{360} & y_i < y_n \end{cases}$$

Pose-based Action Detection



Architecture(channels)	JHMDB-1
(64, 128)	60.11 ± 2.81
(128, 256)	$\mathbf{62.29 \pm 2.50}$
(64, 128, 256)	60.49 ± 3.93
(128, 256, 512)	61.09 ± 4.08

Upper image: Network used for training RPP.

Lower table: Results on JHMDB-1 with various channels and blocks.

Pose-based Action Detection

Concatenation method	JHMDB-1-GT
Stacked(one pathway)	68.51 ± 4.25
Two pathway	$\mathbf{71.38 \pm 2.13}$

Comparisons of two different concatenation methods.

Run ID	mAP
F_M_E_E_BUPT_MCPRL_2	11.6
F_M_E_E_BUPT_MCPRL_1	**11.9**

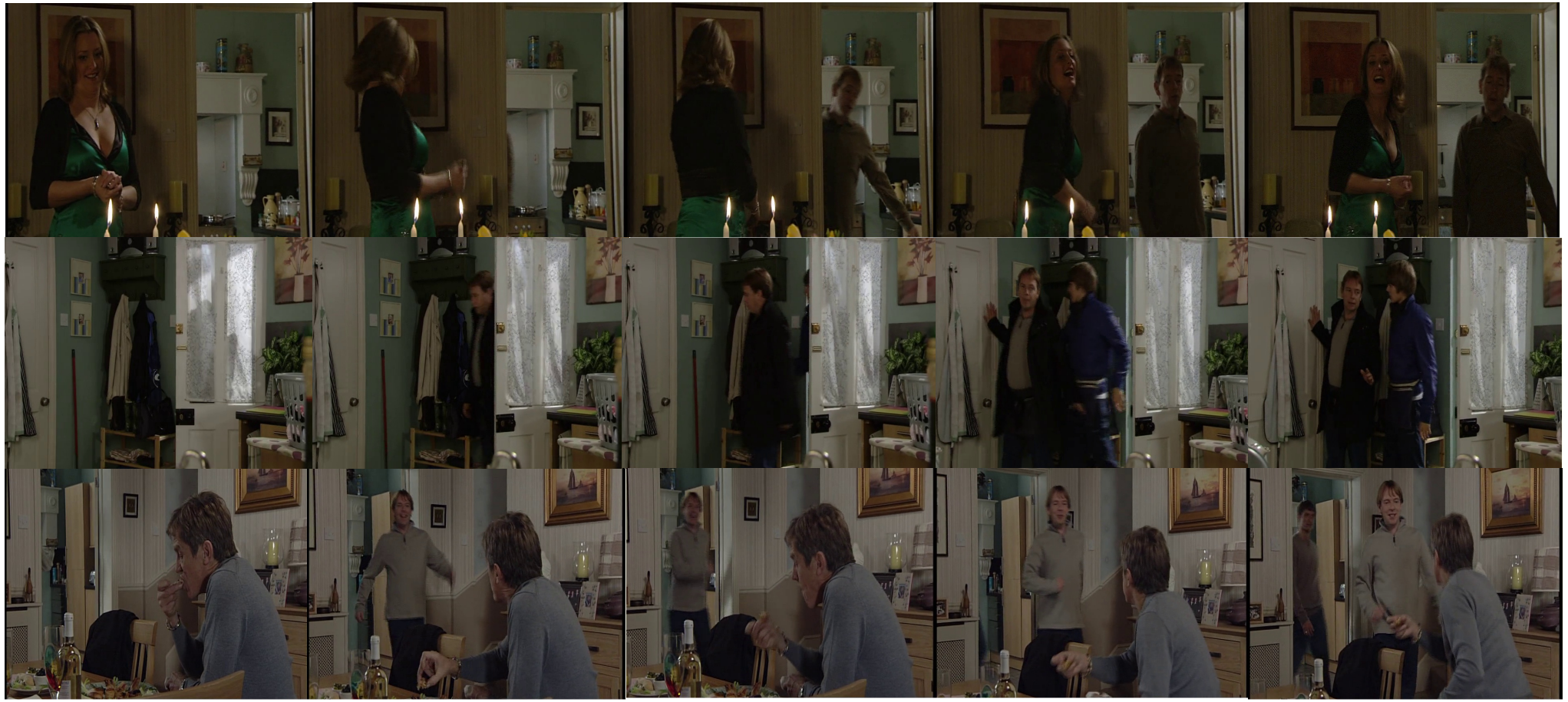
Improvement on INS19.

Methods	JHMDB-1	JHMDB-1-GT
Choutas et. al.	59.1	70.8
Ludl et. al.	60.3 ± 1.3	65.5 ± 2.8
RPR(ours)	$\mathbf{62.29 \pm 2.50}$	$\mathbf{71.38 \pm 2.13}$

Results on JHMDB-1 compared with two state-of-the- art algorithms. JHMDB-1-GT means using pose data given by JHMDB dataset to classify pose representations.

Pose-based Action Detection

Ian Open_door_enter



Conclusion

- Parse INS into several related subtasks and propose a multi-task retrieval framework.
- Detect specific person based on face matching
- Apply expression recognition on related instances
- The semantic dependences of target persons and the corresponding objects are measured to detect human-object interactions
- A light pose-based action detection network and two-pathway ECO are constructed to re-rank INS result list
- The experimental results on four datasets demonstrate the effectiveness of this INS framework

Future work

- Human track
- End-to-end trainable HOI models
- Action localization
- Integrating text and audio information
- More reasonable fusion methods
- …

Thanks!