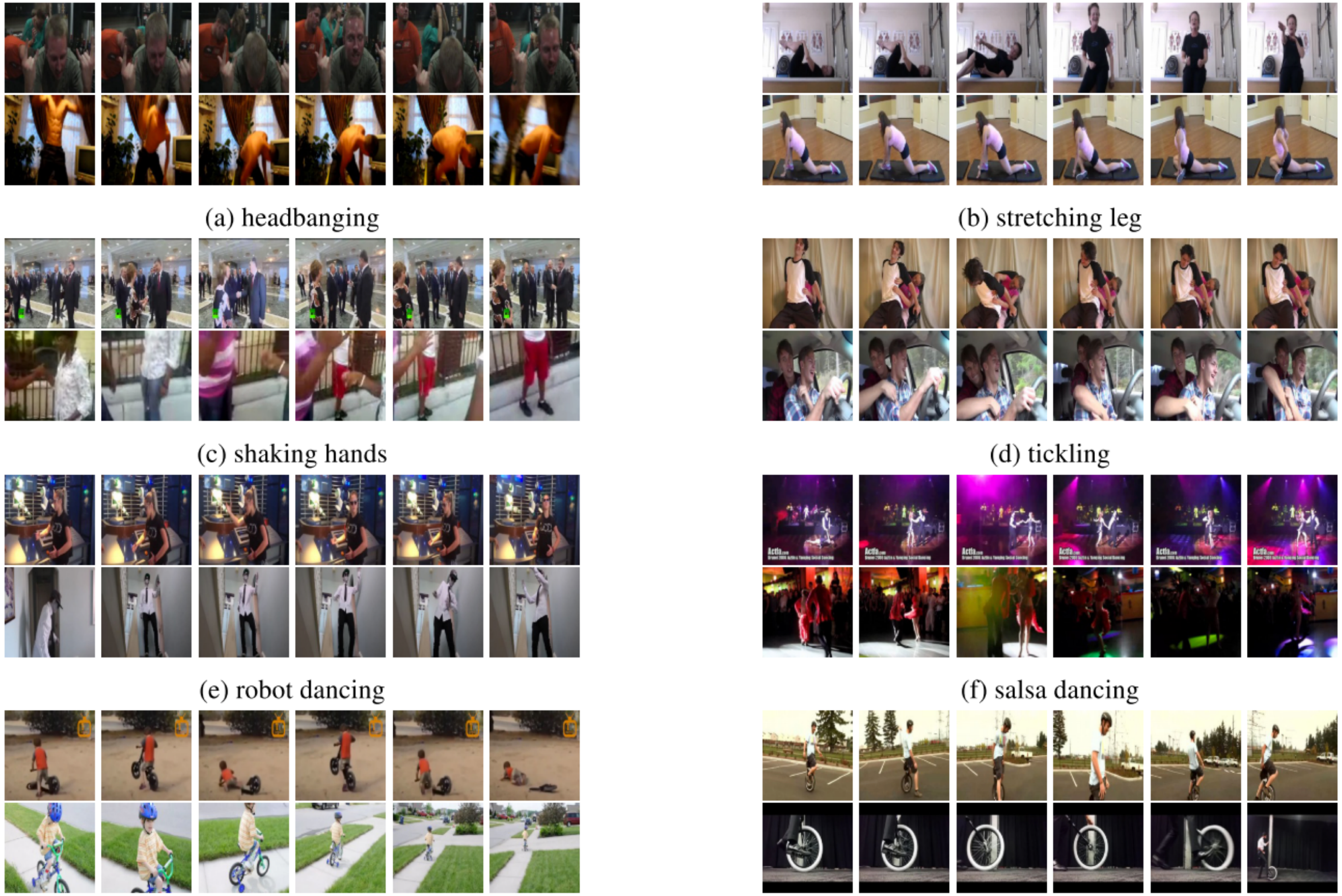
Selecting Informative Video Frames and Optical Flows for Action Recognition with Partial Observations

Yanjun Zhu, Gang Yu, Junsong Yuan, Kai-Kuang Ma

Oct. 9, 2018

What is action recognition?



(a) headbanging

(b) stretching leg

(c) shaking hands

(d) tickling

(e) robot dancing

(f) salsa dancing

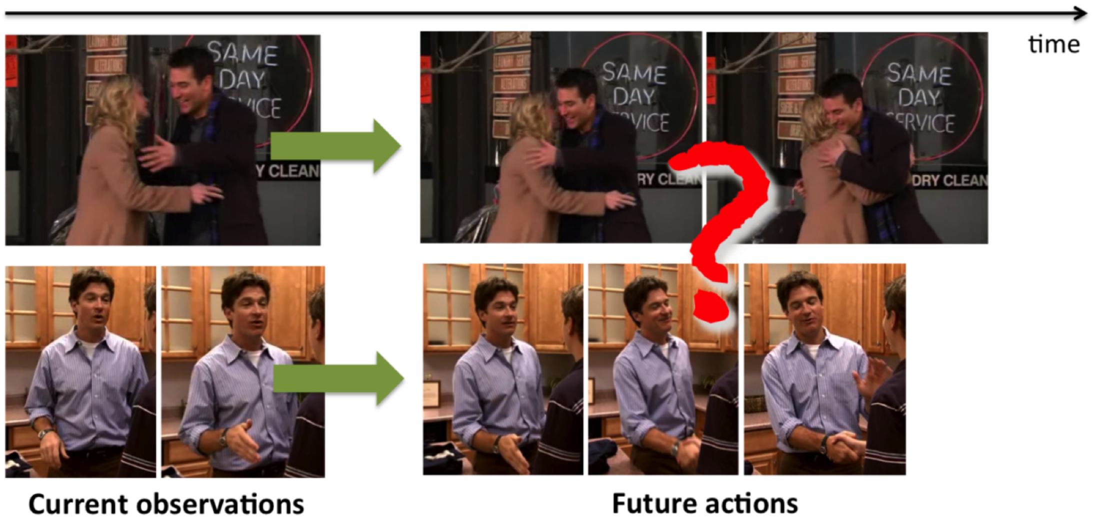
(g) riding a bike

(h) riding unicycle

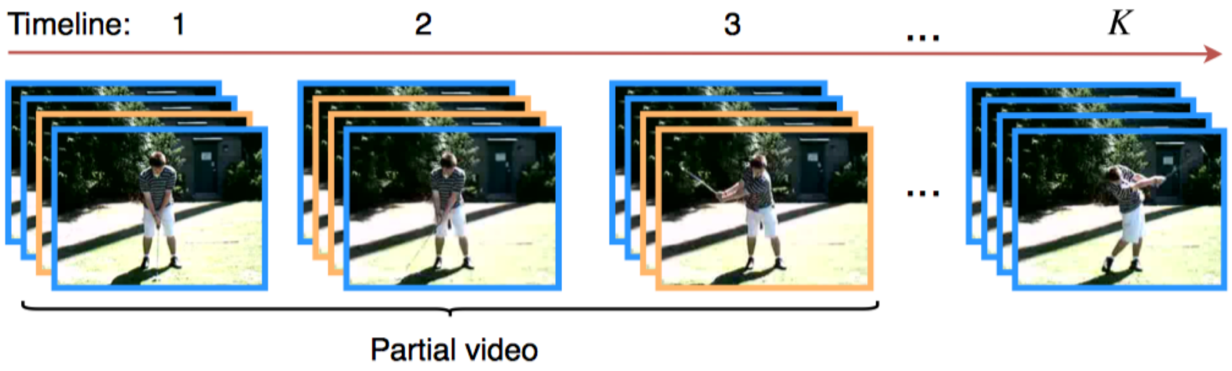
- Recognize action category in trimmed video clips
- Use all video frames
- Give the result after the action completes

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., … Zisserman, A. (2017). The Kinetics Human Action Video Dataset. In *arXiv*. Retrieved from http://arxiv.org/abs/1705.06950

What is partial observations? Why?



Lan, T., Chen, T. C., & Savarese, S. (2014). A hierarchical representation for future action prediction. In *ECCV* (pp. 689–704). https://doi.org/10.1007/978-3-319-10578-9_45



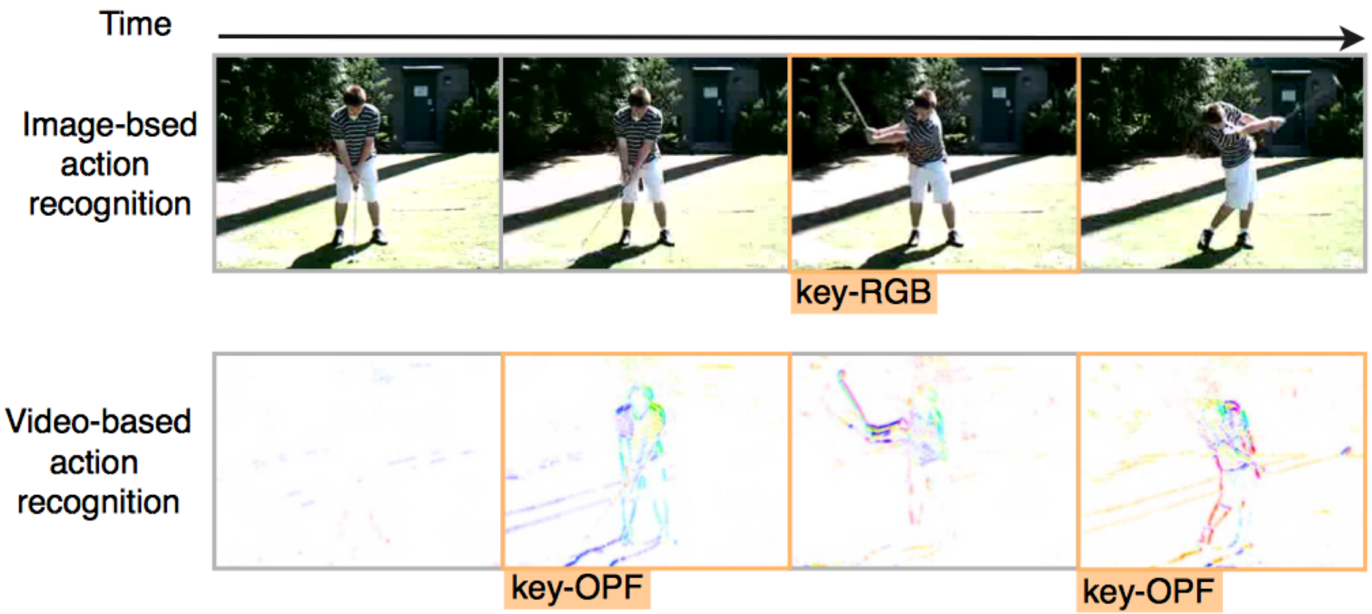
- Give the label using only a part of the video
- Only process the leading frames
- Make a decision before the action completes

Quick response

Why RGB frame and optical flow together?

- RGB frame stands for appearance info
  - Single image based approach
- Optical flow captures motion info
  - Video based approach
  - Stacked optical flow
  - Short-term motion info.

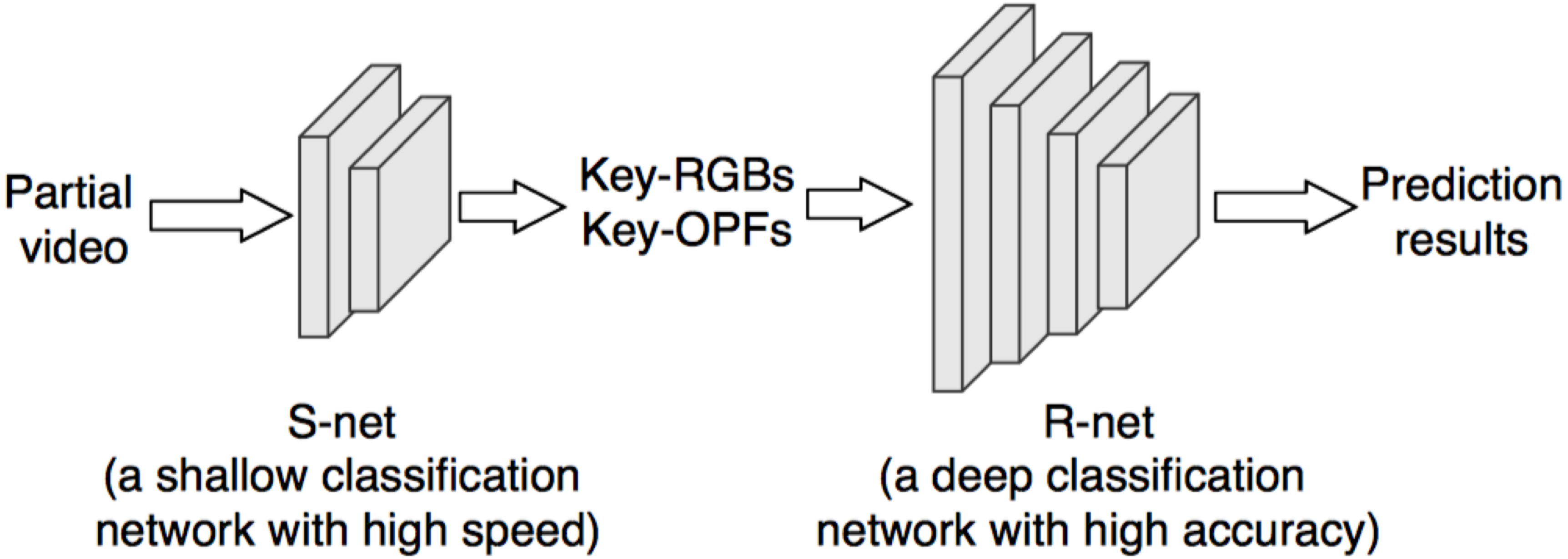
Complementary



Our Approach to Address This Problem

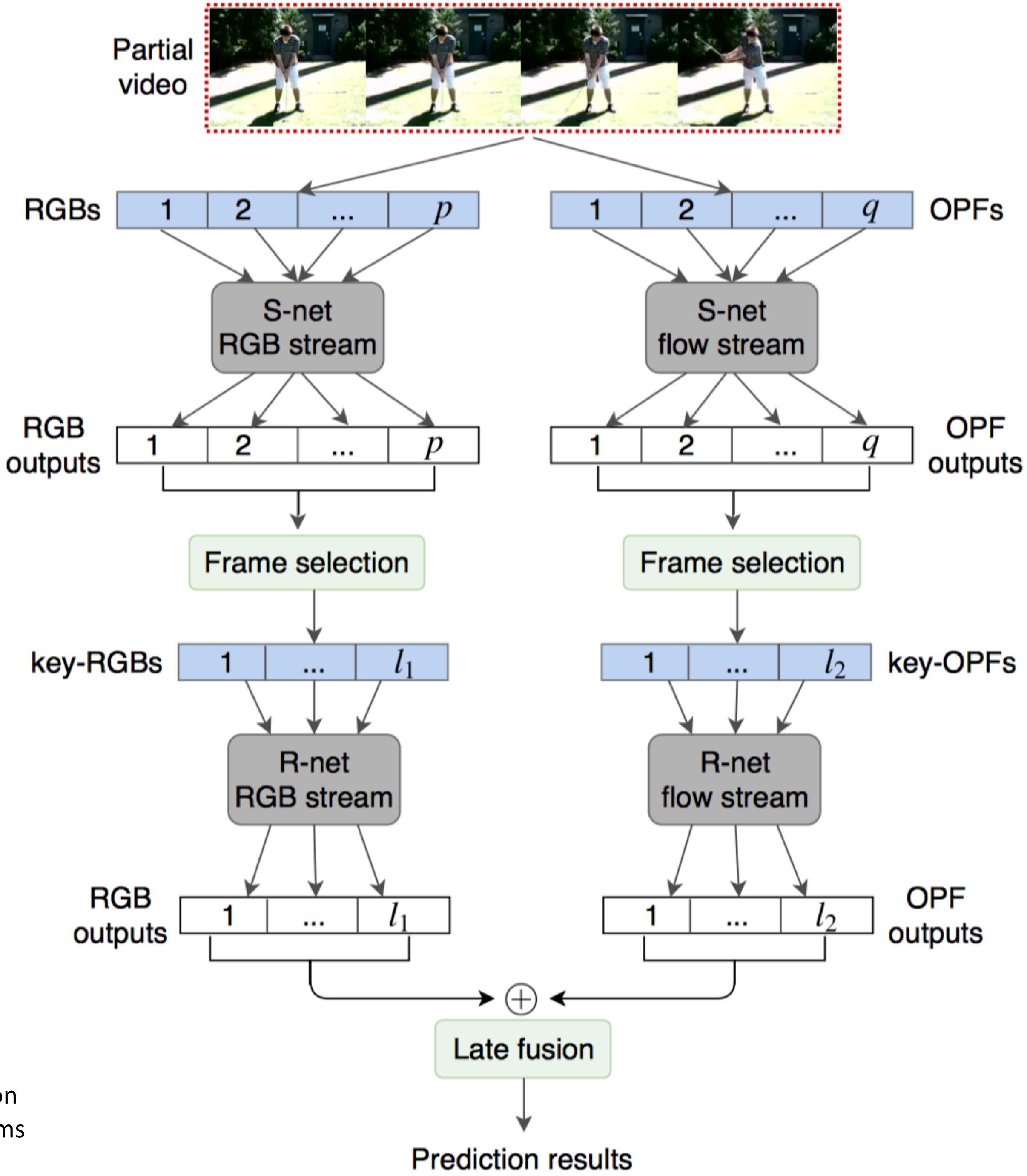
- Select informative frames and optical flows.
- Make decisions with the selected ones.

How?


Partial video
Key-RGBs
Key-OPFs
Prediction results
S-net
(a shallow classification network with high speed)
R-net
(a deep classification network with high accuracy)

Architecture

- Both R-net and S-net follow the two-stream architecture[1]
- The key-RGBs and key-OPFs are selected according to the outputs of the S-net individually.
- Combined together with late fusion (vote)



[1] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. Advances in Neural Information Processing Systems (NIPS), 2014, pp. 568–576.

S-net: Learning to Select Informative Frames

- A shallow classification network with high speed.
- S-net is designed to select key frames to represent the partial video.
- Given a video: $\mathbf{V}$
- The number of pre-defined actions classes: C
- RGB frame: $I_i \in \mathbf{V}, i \in [0, m]$
- Optical flow: $O = \{F_1, \ldots, F_m\}$, where $F_t = \mathcal{G}(I_{t-1}, I_t)$
- S-net outputs a classification score vector: $S_v = \{S_1, \ldots, S_C\}$
- Two criteria to measure the quality of one frame/optical flow:
  - Highest classification score: $S_{max} = \max\{S_1, \ldots, S_C\}$
  - Variance of classification scores: $S_{var} = \mathrm{var}\{S_1, \ldots, S_C\}$

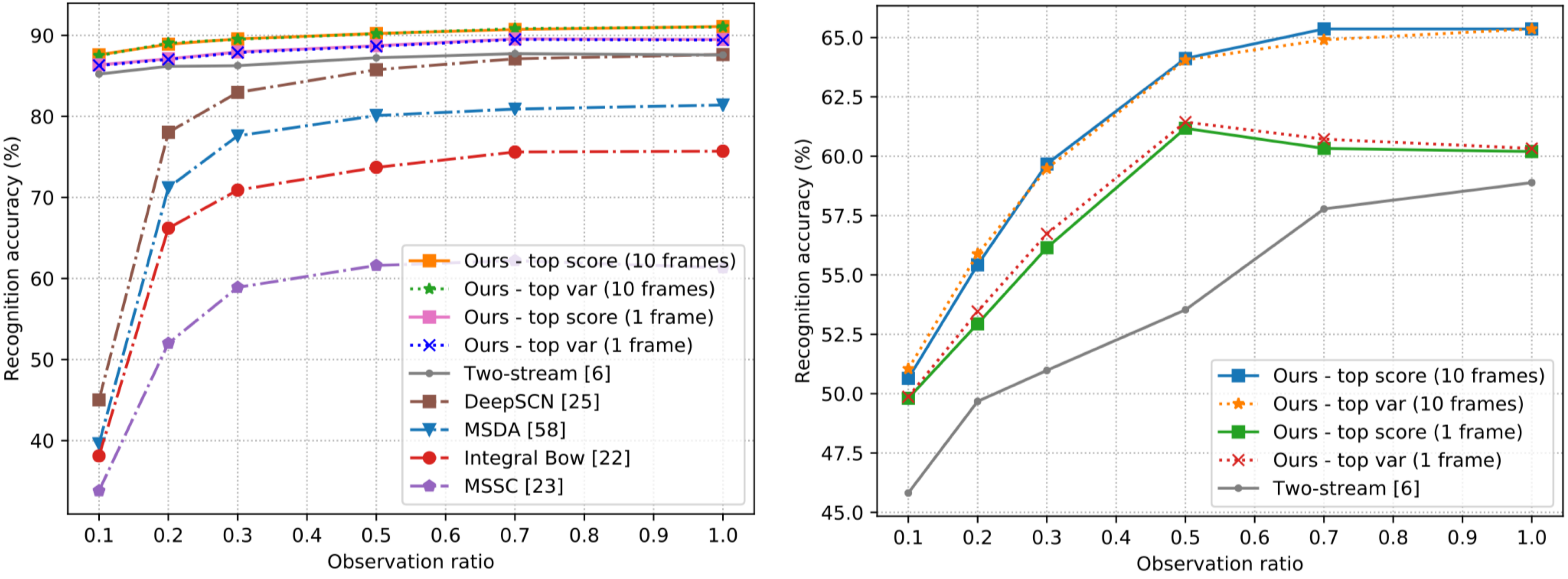
R-net: Recognizing Actions with Informative Frames and Optical Flows

- A deep classification network with high accuracy.
- Each stream takes one frame as input each time, and the final recognition result is determined by the outputs of multiple key-RGB and key-OPF.
- Two step fusion:
  - Internal: average score vectors of all key frames in each stream
  - External: weighted average of the two streams

Training Strategy

- All the frames of the video dataset are used as training samples.
- R-net and S-net are trained separately
- The RGB stream and the flow stream of them are also trained separately
- Cross entropy loss
- Optimizer: SGD with momentum
- The optical flow is pre-computed using TVL1 algorithm

Experiments: Recognition Accuracy

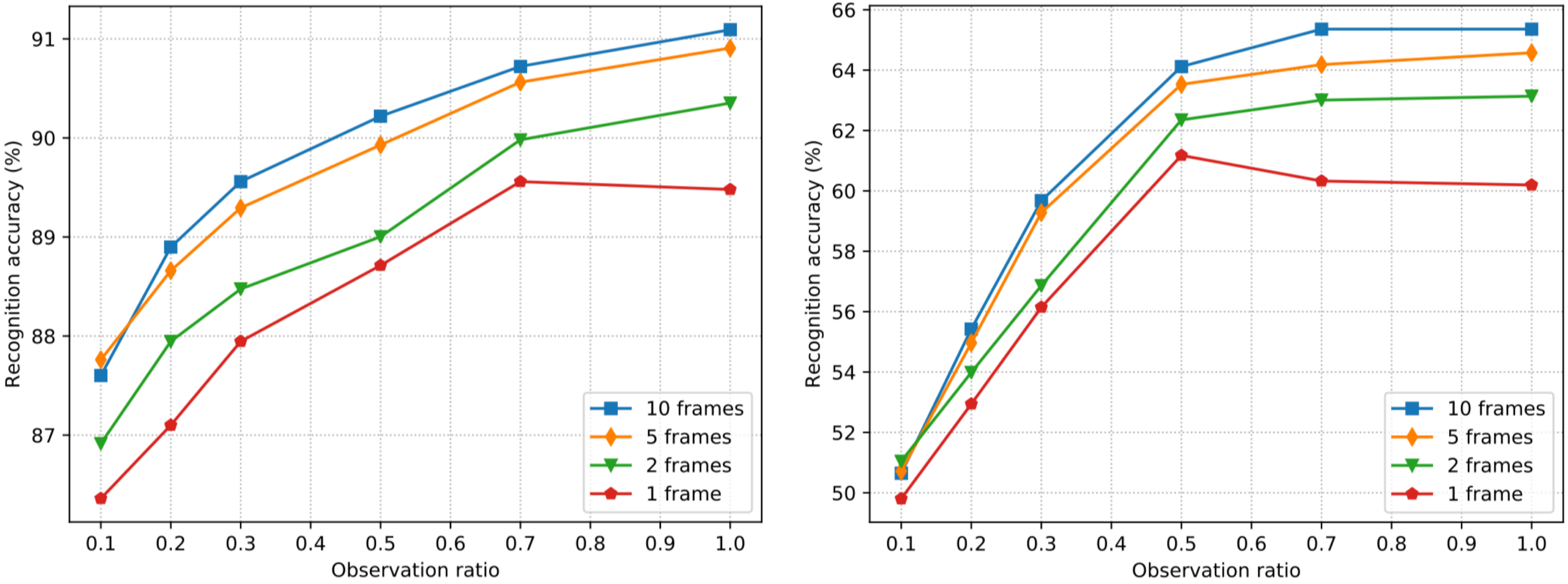


(a) UCF101

(b) HMDB51

Fig. 4. The results of action recognition with partial observations on UCF101 and HMDB51. The original two-stream method, i.e., R-net with uniform selection is considered as the comparison baseline. The recognition accuracy of both top score and top variance selection is listed. Ten key frames and 1 key frame in each partial video are used in the two selection methods.

Experiments: Number of Key Frames



(a) UCF101

(b) HMDB51

Fig. 6. Evaluations on different number of key frames. In each partial video, {1, 2, 5, 10} key frames are selected by the S-net respectively. Then they are fed into the R-net for final recognition and the recognition accuracy of every observation ratio is shown.

Experiments: Frame Selection Criteria

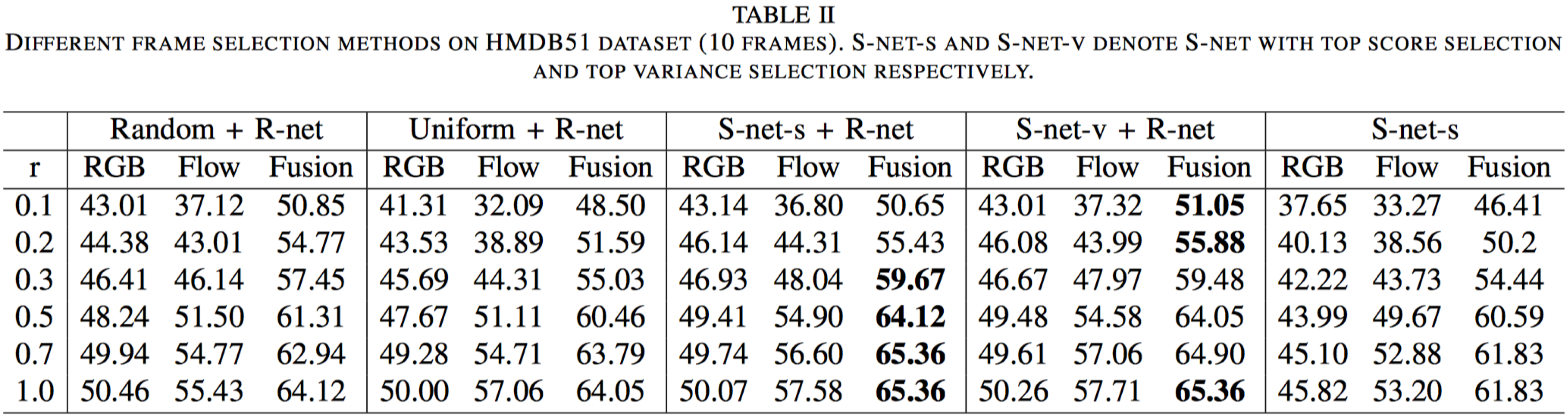


TABLE II
DIFFERENT FRAME SELECTION METHODS ON HMDB51 DATASET (10 FRAMES). S-NET-S AND S-NET-V DENOTE S-NET WITH TOP SCORE SELECTION AND TOP VARIANCE SELECTION RESPECTIVELY.

	Random + R-net			Uniform + R-net			S-net-s + R-net			S-net-v + R-net			S-net-s		
r	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion
0.1	43.01	37.12	50.85	41.31	32.09	48.50	43.14	36.80	50.65	43.01	37.32	**51.05**	37.65	33.27	46.41
0.2	44.38	43.01	54.77	43.53	38.89	51.59	46.14	44.31	55.43	46.08	43.99	**55.88**	40.13	38.56	50.2
0.3	46.41	46.14	57.45	45.69	44.31	55.03	46.93	48.04	**59.67**	46.67	47.97	59.48	42.22	43.73	54.44
0.5	48.24	51.50	61.31	47.67	51.11	60.46	49.41	54.90	**64.12**	49.48	54.58	64.05	43.99	49.67	60.59
0.7	49.94	54.77	62.94	49.28	54.71	63.79	49.74	56.60	**65.36**	49.61	57.06	64.90	45.10	52.88	61.83
1.0	50.46	55.43	64.12	50.00	57.06	64.05	50.07	57.58	**65.36**	50.26	57.71	**65.36**	45.82	53.20	61.83

Experiments: Speed

- Intel Xeon CPU @2.20GHz, NVIDIA Titan Xp GPU

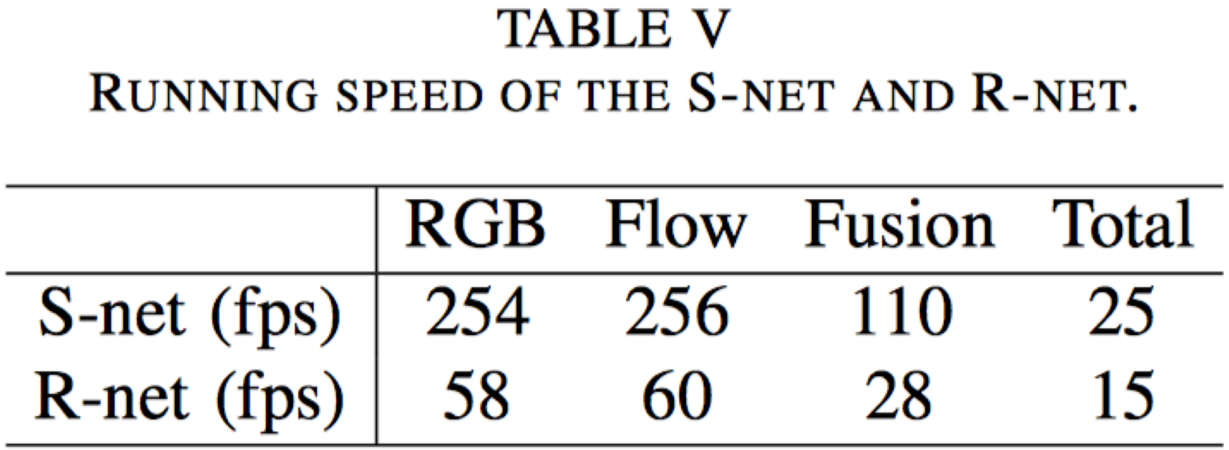


TABLE V
RUNNING SPEED OF THE S-NET AND R-NET.

	RGB	Flow	Fusion	Total
S-net (fps)	254	256	110	25
R-net (fps)	58	60	28	15

Experiments: Test on Streaming Videos

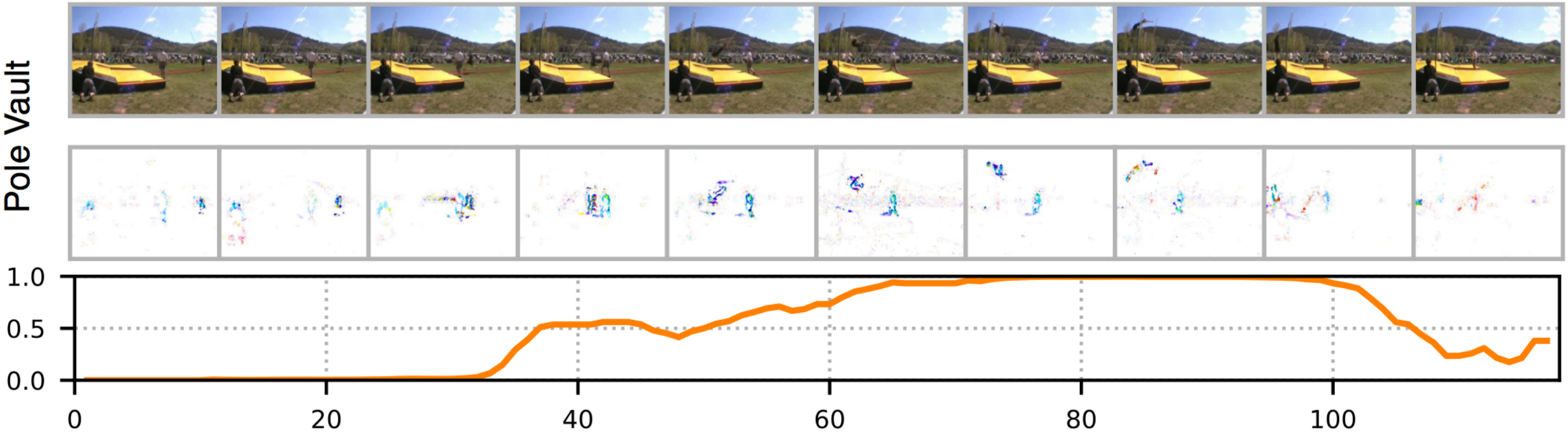
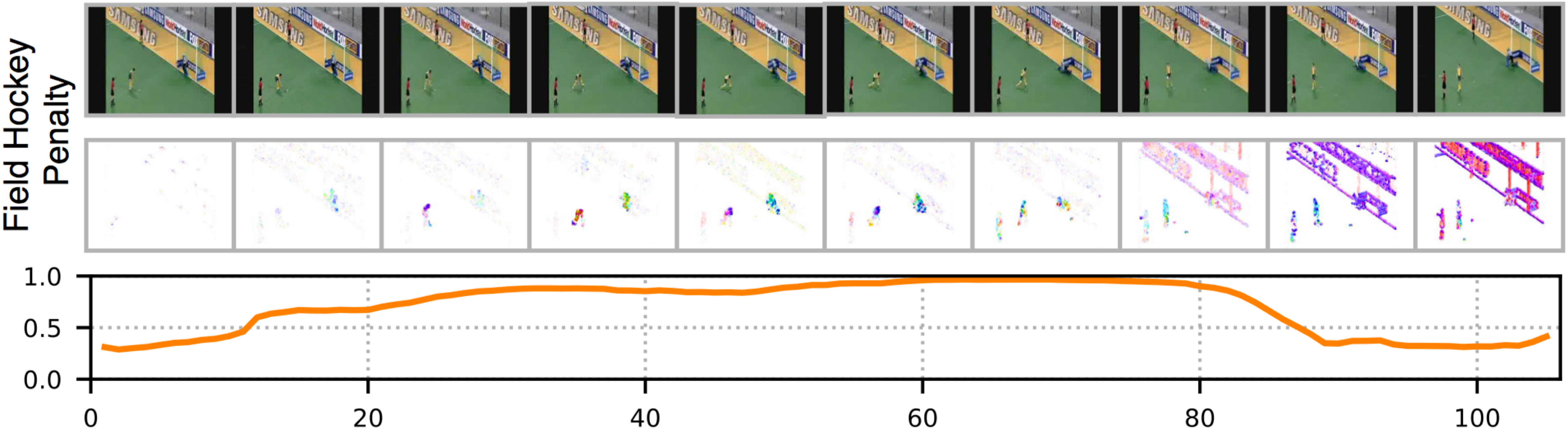
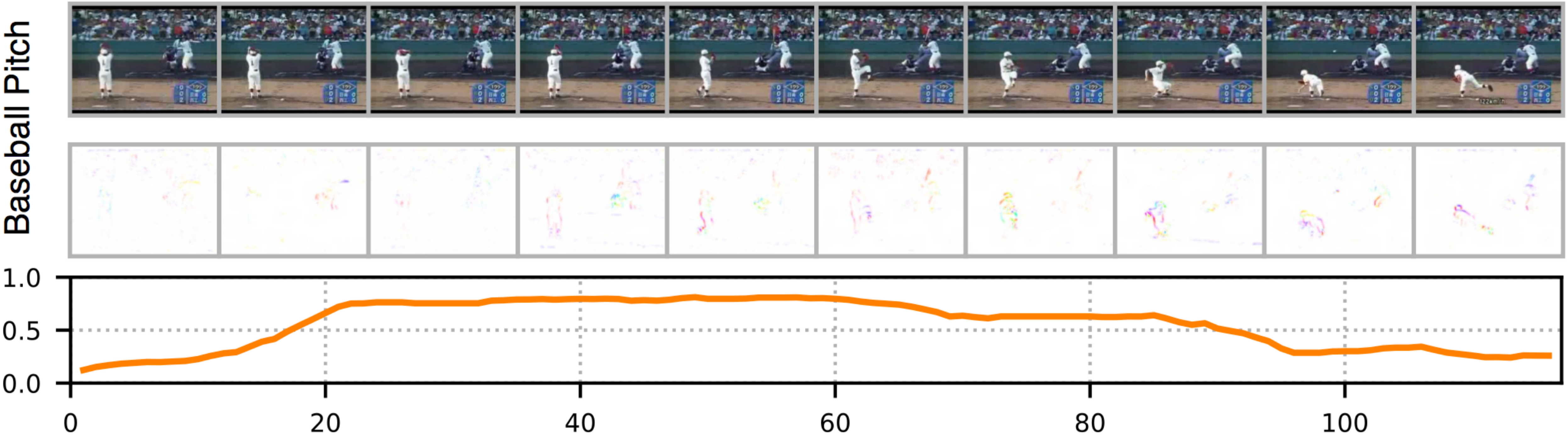


Fig. 8. The curves of action recognition scores on streaming videos. Three actions are listed here, namely field hockey penalty, baseball pitch and pole vault. All the videos are processed one frame by one frame along the time line. In each action video, RGB frames and optical flows are listed in the first and second rows respectively. The orange curve in the third row indicates the recognition score of current observations. The higher the score is, the more confidently that the current observations are predicted to its true class. The horizontal axis of the curve indicates the number of frames along the timeline.


Field Hockey
Penalty
1.0
0.5
0.0
0
20
40
60
80
100


Baseball Pitch
1.0
0.5
0.0
0
20
40
60
80
100

Thank you!

Any questions, please contact the first author Yanjun Zhu.

yanjun_zhu@outlook.com