Residual Stacked RNNs for Action Recognition

Mohamed Ilyes Lakhal*, Albert Clapés, Sergio Escalera, Oswald Lanz, Andrea Cavallaro

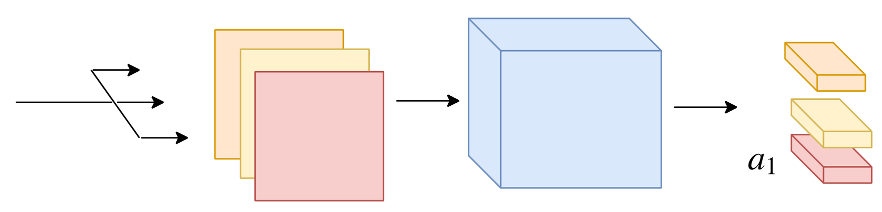
m.i.lakhal@qmul.ac.uk

1. Introduction

- Getting the best from both successful deep networks:
  - 3D-CNN for spatio-temporal feature extraction
  - RNN as a temporal dependencies learning.
- Do we benefit from residual learning in RNN as CNN models do?
- *Solution:* a two-stream stacked residual RNN

2. Stacked residual RNN

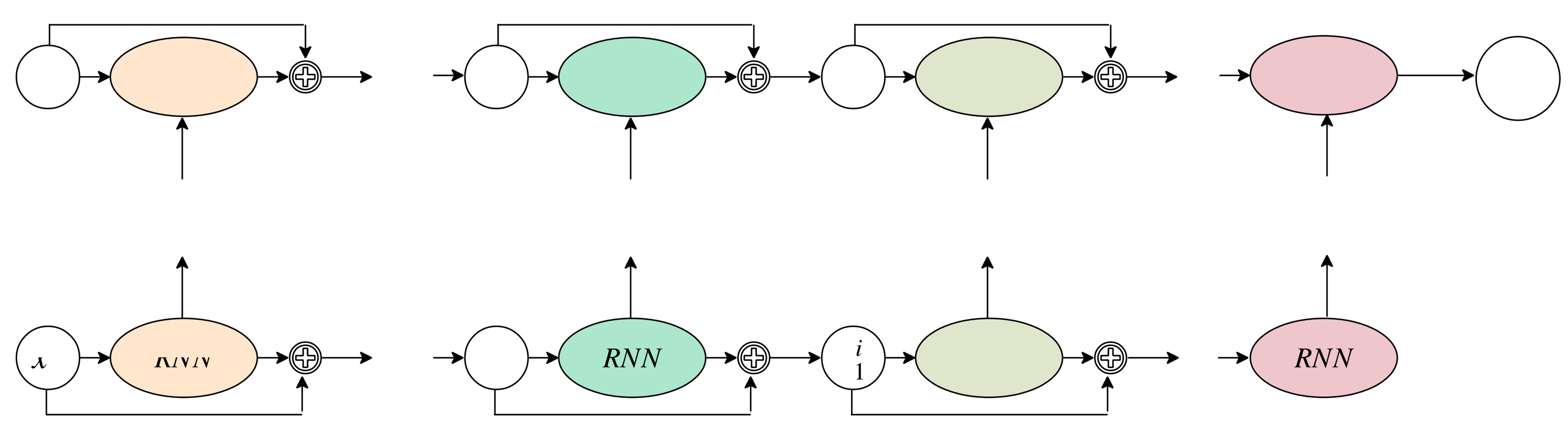
Spatio-temporal feature extraction



- We divide an input video $V \in \mathbb{R}^{m_x \times m_y \times l_v}$ into K_v segments
- We feed each segment S_i into 3D-CNN to extract spatio-temporal feature: $a_i = E(S_i) \in \mathbb{R}^{d_f}$

$$a = \{a_i | i = 1 \dots K_v; a_i \in \mathbb{R}^{d_f}\}$$

Sequence learning

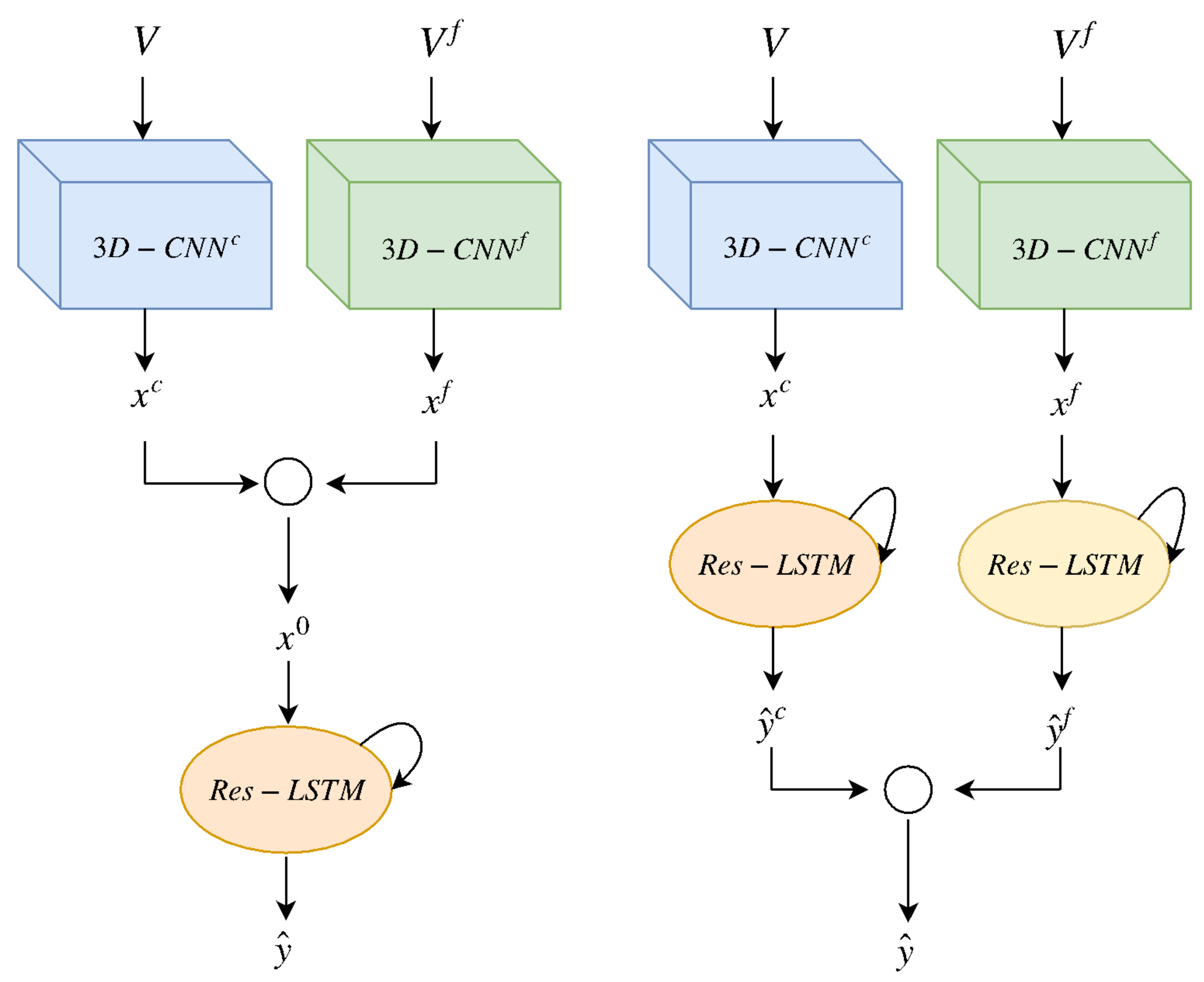


- The input to the Residual LSTM is given by:

$$x^0 = \{x_t^0 = a_{\sigma(t)} | t = 1, \dots, T; a_{\sigma(t)} \in a\}$$

- The class label prediction is obtained at time step *T* of the L^{th} LSTM model

Two stream model



Mid-fusion

Late-fusion

The feature (or score) aggregation is performed using:

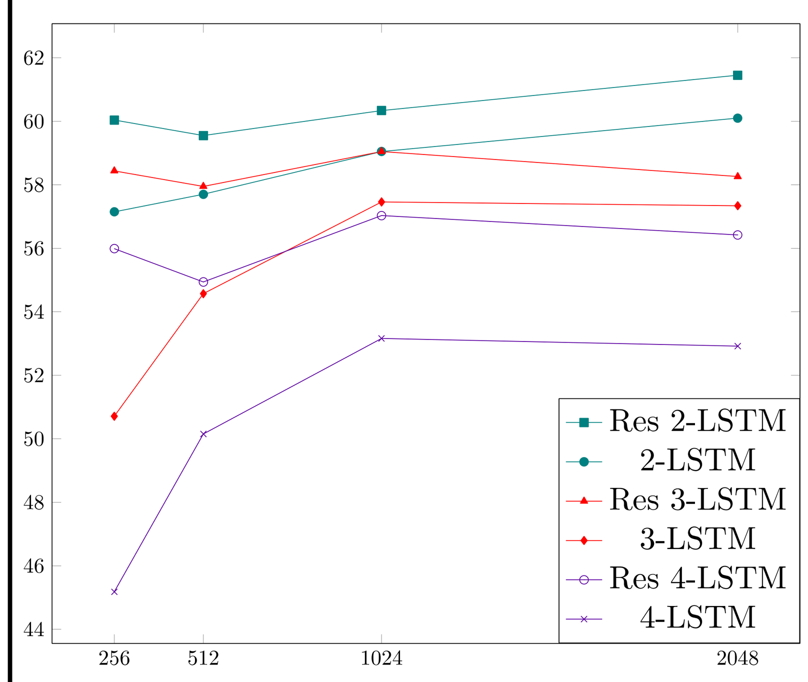
Element-wise sum: $u \oplus v = (u_1 + v_1, \dots, u_m + v_m)$

Element-wise product: $u \odot v = (u_1.v_1, \dots, u_m.v_m)$

3. Experimental results

Datasets: UCF-101 [5] and HMDB-51 [6]

Model selection



LSTM vs Res-LSTM (HMDB-51)

Strategy	Mid fusion	Late fusion
$\oplus$ (element-wise sum)	59.3	65.2
$\odot$ (element-wise product)	56.5	68.0
$w_1.Flow + w_2.RGB$	–	63.3

T	5	15	25	35
HMDB-51	59.5	60.2	61.5	61.4
UCF-101	77.9	79.9	79.5	80.9

State-of-the-art comparison

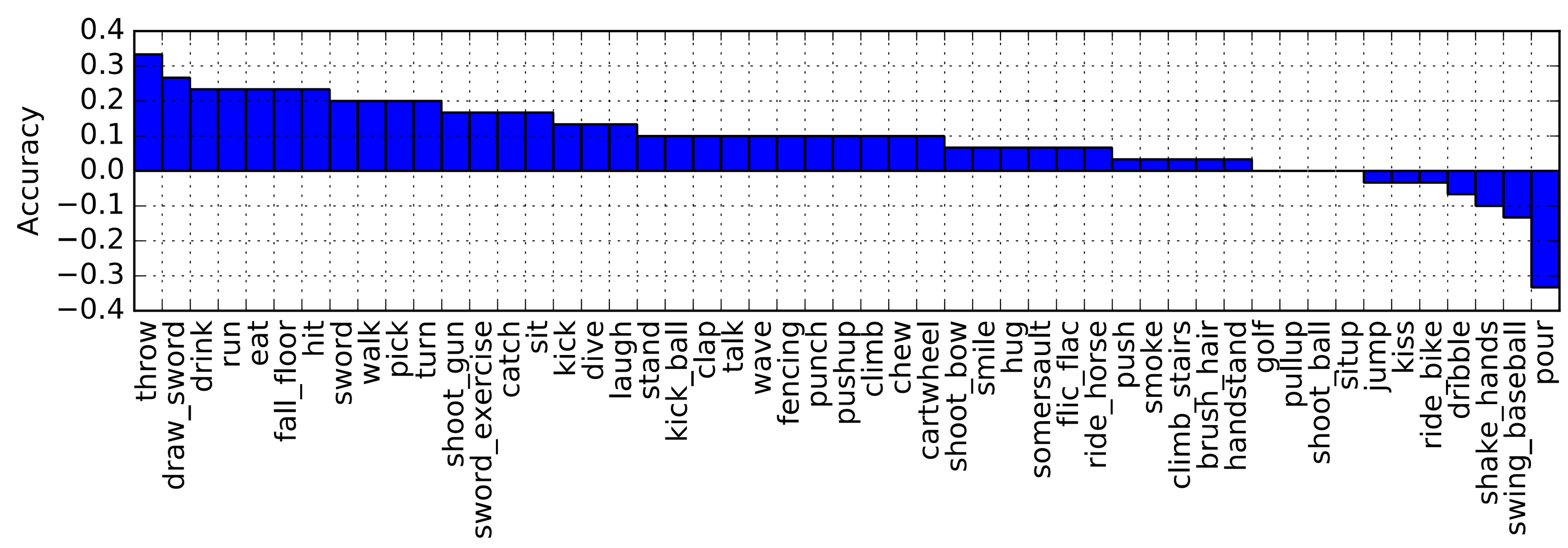
Model	Method	UCF-101	HMDB-51
Static	STC-ResNext [1]	95.8‡	72.6‡
	I3D [2]	**98.0**	**80.7**
Sequential	L^2STM [3]	**93.6**	66.2
	PreRNN [4]	**93.7***	–
	Res-LSTM (ours)	92.5*	68.0*
	Res-LSTM (ours) $\odot$ IDT	93.0*	**76.9***

‡ Only RGB modality is used

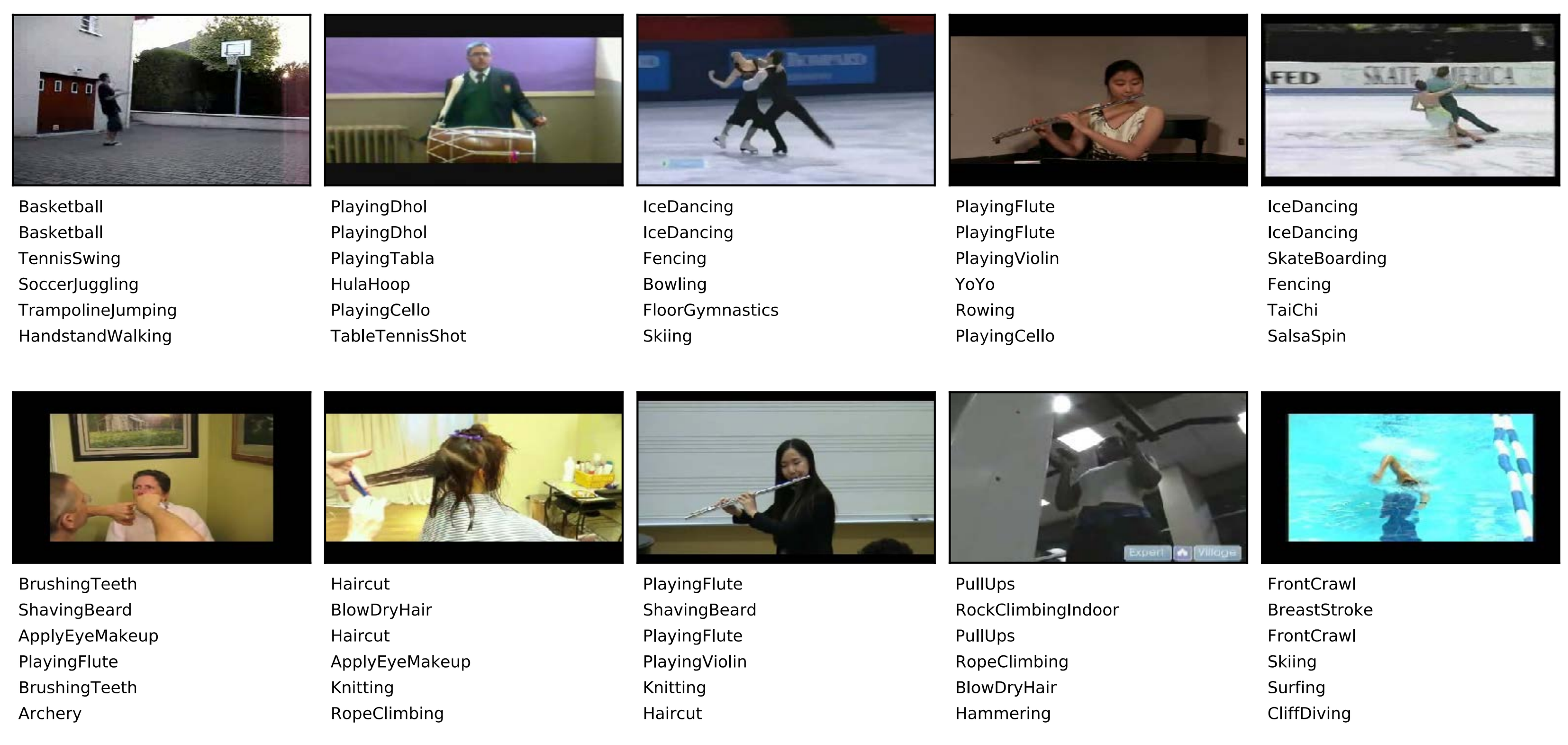
* Evaluation on split-1

Data Types	L^2STM	Res-LSTM	Gain
Human-Object Interaction	86.7	88.2	↑ **1.5**
Human-Human Interaction	95.4	96.9	↑ **1.5**
Body-Motion Only	88.6	90.7	↑ **2.1**
Playing Instrument	-	97.3	-
Sports	-	93.2	-
Pizza Tossing	72.7	66.7	↓ 6.0
Mixing Batter	86.7	91.1	↑ **4.4**
Playing Dhol	100	100	≡
Salsa Spins	100	97.7	↓ 2.3
Ice Dancing	100	100	≡

Per-class accuracy improvement when combining with IDT [7]



Top-5 class predictions



4. Conclusions

- Proposed a two-stream stacked Res-LSTM to action recognition by means of spatio-tempoal feature instead of frame based approach
- Our proposed model shows state-of-the-art results on HMDB-51 for RNN-like solutions and good performance overall.

References

[1] Ali, D., Mohsen, F., Vivek, S., M.Mahdi, A., Rahman, Y., Juergen, G., Van Gool, L.: Spatio-temporal channel correlation networks for action classification. ECCV 2018.

[2] Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. CVPR 2017.

[3] Sun, L., Jia, K., Chen, K., Yeung, D.Y., Shi, B.E., Savarese, S.: Lattice long shortterm memory for human action recognition. ICCV 2017

[4] Yang, X., Molchanov, P., Kautz, J.: Making convolutional networks recurrent for visual sequence learning. CVPR 2018

[5] Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR 2012

[6] Kuehne, H., Jhuang, H.h., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. ICCV 2011

[7] H. Wang and C. Schmid. Action recognition with improved trajectories. ICCV, 2013

Acknowledgment

This work has been partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.