Maryam Asadi-Aghbolaghi[1,2,3], Albert Clapés[2,3], Marco Bellantonio[4], Hugo Jair Escalante[5], Victor Ponce-López[2,3], Xavier Baró[3,6], Isabelle Guyon[7], Shorheh Kasaei[1], Sergio Escalera[2,3]
[1]Dept. of Computer Engineering, Sharif University of Technology, Tehran, Iran | [2]Dept. of Applied Mathematics and Analysis, Univessity of Barcelona, Barcelona, Spain | [3]Comuter Vision Center, Autonomous University of Barcelona, Bellaterra (Barcelona), Spain | [4]School of Informatics, Polytechnic University of Barcelona, Barcelona, Spain | [5]Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México | [6]EIMT, Open University of Catalonia, Barcelona, Spain | [7]Université Paris-Saclay, Paris, France

A survey on deep learning based approaches for action and gesture recognition in image sequences

Abstract

In this paper, we present:

- A **survey** on current deep learning methodologies for action and gesture recognition in image sequences,
- A **taxonomy** that summarizes important aspects of deep learning for approaching both tasks with particular interest on how they treat the temporal dimension of data,
- The details of the proposed **architectures**, **fusion** strategies, main **datasets**, and **competitions**,

Motivation

The interest in **action and gesture recognition** has grown considerably in the last years. Recent deep learning outperformed "non-deep" state-of-the-art methods. However, some questions remain opened:

- How to deal with **temporal information**. We investigated works that go beyond averaging class score predictions on individual frames for video prediction.
- How to be train deep models with **small datasets.**
- Whether deep-learning approaches rely only on deep models or in **combination with hand-crafted features.**
- Which are the most successful approaches **to anticipate future trends and research directions.**

Datasets and challenges

Action datasets

Year	Dataset	Problem	Body Parts	Modality	No.classes	Performance
2008	UCF Sports	AC, STL	F	RGB	10	95.80%, 0.789@0.5 mAP
2009	Hollywood 2	AC	F, U, L	RGB	12	78.50 mAP
2010	Highfive	AC, STL	F,U	RGB	4	69.40 mAP [7], 0.466 IoU
2010	Olympic Sports	AC	F	RGB	16	96.60% Acc [6]
2011	HMDB51	AC	F, U, L	RGB	51	73.60% Acc
2012	MPII Cooking	AC, TL	F, U	RGB	65	72.40 mAP, -
2012	UCF101	AC,TL	F, U, L	RGB	101	94.20% Acc [13], 46.77@0.2 mAP (split 1)
2014	Sports 1-Million	AC	F, U, L	RGB	487	73.10% Acc
2014	THUMOS-14	AC, TL	F, U, L	RGB	101, 20 *	71.60 mAP [8], 0.190@0.5 mAP [11]
2015	THUMOS-15	AC, TL	F, U, L	RGB	101, 20 *	80.80 mAP [6], 0.183@0.5 mAP
2015	ActivityNet	AC, TL	F, U, L	RGB	200	93.23 mAP, 0.594@0.5 mAP

Gesture datasets

Year	Dataset	Problem	Body Parts	Modality	No.classes	Performance
2011	ChaLearn Gesture	GC	F, U	RGB, D	15	-
2012	MSR-Gesture3D	GC	F, H	RGB, D	12	98.50% Acc
2014	ChaLearn (Track 3)	GC, TL	U	RGB, D, S	20	98.20 Acc, 0.870 IoU
2015	VIVA Hand Gesture	GC	H	RGB	19	77.50% Acc
2016	ChaLearn conGD	TL	U	RGB, D	249	0.315 IoU
	ChaLearn isoGD	GC				67.19% Acc

Challenges

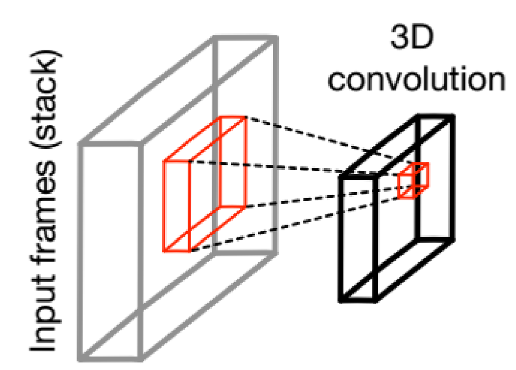
Challenge	Year	Dataset	Task	Event
ChaLearn	2012	CGD	G	-
	2013	Montalbano	G	-
	2014	HuPBA 8K+	A	ECCV
		Montalbano	G	
	2015	HuPBA 8K+	A	CVPR
	2016	isoGD, conGD	G	ICPR
HAL	2012	LIRIS	A	ICPR
Opportunity	2011	Opportunity	A	-
ROSE	2016	NTU RGB+D	A	ACCV
THUMOS	2013	UCF101	A	ICCV
	2014	THUMOS-14	A	ECCV
	2015	THUMOS-15	A	CVPR
VIVA	2015	VIVA	G	CVPR
VIRAT	2012	VIRAT DB	A	CVPR

Taxonomy

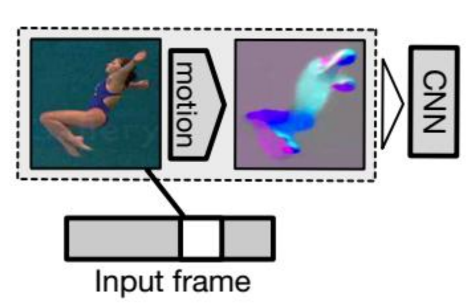
Architectures

We categorize the different CNN-based approaches based on how they **handle the temporal dimension of videos**:

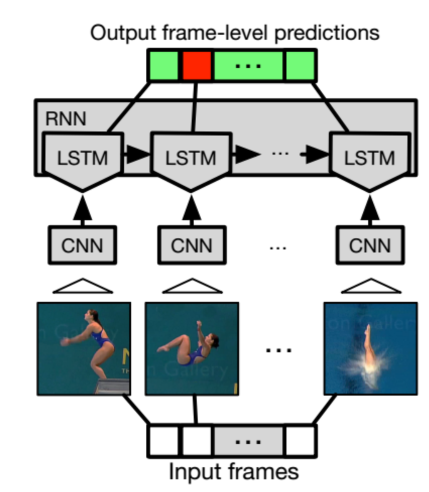
- **3D convolutions** which are able to learn local spatiotemporal features by extending the connectivity of convolutional neurons across multiple adjacent frames [1].
- **Motion hand-crafted features** (e.g. dense optical flow frames) being directly input a second cue along with the color one [2,4,7].
- **Sequential models** (e.g. CNN+RNN) that model the evolution of responses got from a frame-level classifier [12].



3D convolutions



Motion-based hand-crafted features



Sequential models

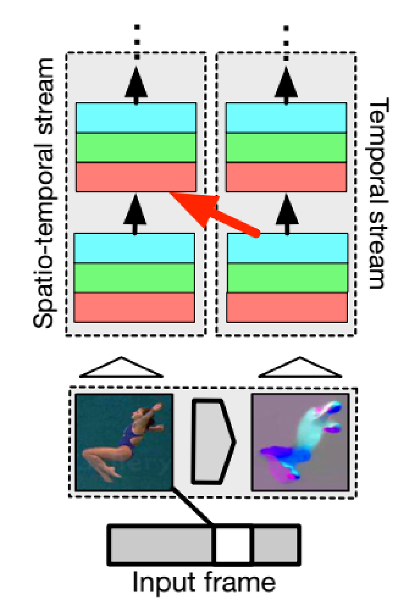
Fusion strategies

The goal is to exploit information complementariness and redundancy for improving the recognition performance, either by using:

- Several **frames, fixed-length clips, or spatial locations** sampled across the entire video.
- **Multiple data cues** (color, motion, depth, etc).

The most **common strategies** can be categorized into:

- **Early fusion:** stacking the information as different input channels [1].
- **Late fusion:** combining class predictions [4].
- **Slow fusion:** progressively fusing by convolution and pooling.



Slow fusion

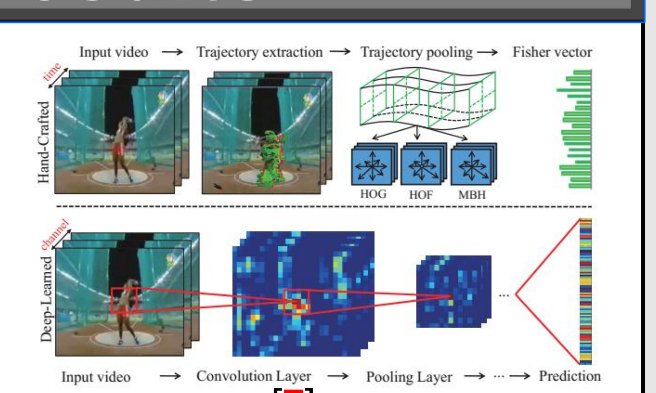
State-of-the-art method results

UCF-101 results

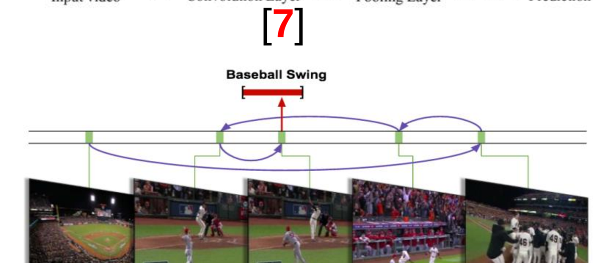
Ref.	Year	Features	Architecture	Score
[2]	2015	CNN, IDT	2 CNN + iDT pooling	93.78%
[3]	2016	Opt. Flow, 3D CNN, IDT	LTC-CNN	92.7%
[4]	2016	conv5, 3D pool	VGG-16, VGG-M, 3D CNN	92.5%
[5]	2016	CNN	Siamese VGG-16	92.4%
[6]	2016	CNN fc7	2 CNNs (spatial + temporal)	92.2%
[7]	2015	CNN, Hog/Hof/Mbh	2-stream CNN	91.5%

THUMOS'14 results

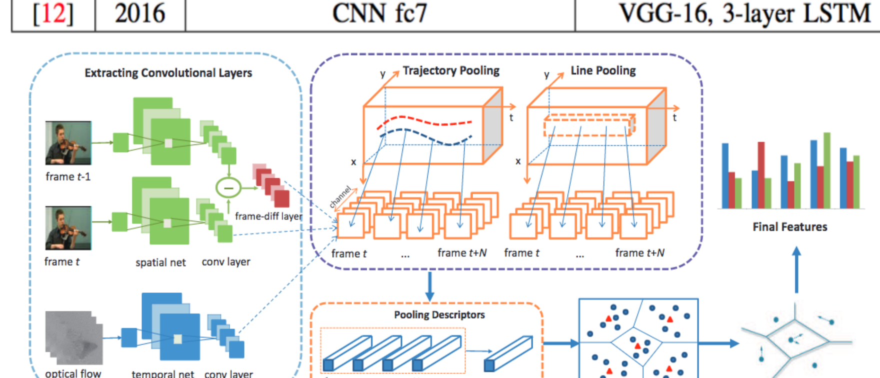
Ref.	Year	Features	Architecture	Score
[8]	2015	H/H/M, IDT, FV+PCA+GMM.	8-layer CNN	71.6%
[9]	2016	CNN	2 CNNs (spatial + temporal)	61.5%
[10]	2015	ImageNet CNN, word2vec GMM	CNN	56.3%
[11]	2016	CNN fc6, fc7, fc8	3D CNN, Segment-CNN	19% mAP
[12]	2016	CNN fc7	VGG-16, 3-layer LSTM	17.1% mAP



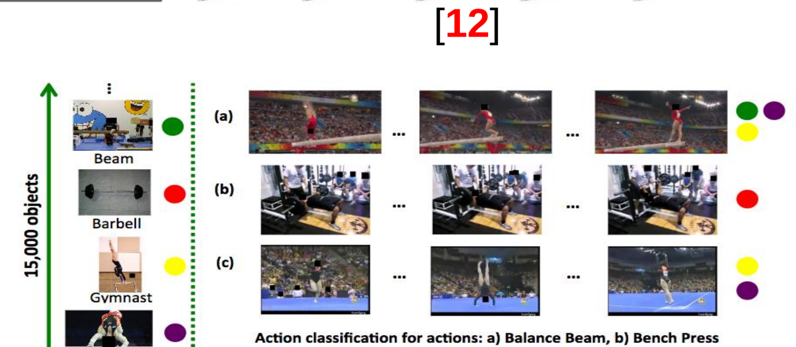
[7]



[12]



[2]



[8]

Discussion

On **temporal modeling**:

- The most input *naive* way to deal with temporal dimension is to average frame-level class score predictions (not covered in this version of the paper).
- 3D convolutions can model discriminative – more local – spatiotemporal features [1].
- Sequential models (e.g. LSTM) better handle longer-range temporal relations.
- It has proven useful to sacrifice spatial resolution in favor of extending the temporal connectivity in the network's input (i.e., larger clips) [3].

On training with **small datasets**:

- Motion (e.g. optical flow) and skeleton features are easier to model (and not *overfit*) than appearance [18].
- When using 2D CNNs, image datasets (e.g. ImageNet) are often used to pre-train weights of the appearance (namely spatial) stream [18].
- 3D CNNs can be initialized using 2D weights [14].
- Multi-task learning has proven useful to jointly train on several datasets (loss function combining several soft-max layers' outputs) [18].

On the exploitation of **hand-crafted features** in hybrid approaches:

- Video class predictions from hand-crafted approaches are combined with the ones from the deep model [7].
- In particular, iDTs can be used to pool deep features from CNN convolutional maps [2,7].
- Taking advantage of human body spatial constraints [16] or interaction among subjects [17].

On **future trends and research directions**:

- Towards more complex end-to-end trained models [12].
- Efficient recognition and detection of actions in more complex longer sequences [11,12].
- Early detection [15].

References

[1] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *TPAMI* 35.1 (2013): 221-231.
[2] Zhao, Shichao, et al. "Pooling the convolutional layers in deep convnets for action recognition." *arXiv preprint arXiv:1511.02126* (2015).
[3] Varol, Gül, Ivan Laptev, and Cordelia Schmid. "Long-term temporal convolutions for action recognition." *arXiv preprint arXiv:1604.04494* (2016).
[4] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[5] Wang, Xiaolong, Ali Farhadi, and Abhinav Gupta. "Actions~transformations." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[6] Li, Yingwei, et al. "Vlad3: Encoding dynamics of deep features for action recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[7] Wang, Limin, Yu Qiao, and Xiaoou Tang. "Action recognition with trajectory-pooled deep-convolutional descriptors." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
[8] Jain, Mihir, Jan C. van Gemert, and Cees GM Snoek. "What do 15,000 object categories tell us about classifying and localizing actions?." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
[9] Zhang, Bowen, et al. "Real-time action recognition with enhanced motion vector CNNs." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[10] Jain, Mihir, et al. "Objects2action: Classifying and localizing actions without any video example." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.
[11] Shou, Zheng, Dongang Wang, and Shih-Fu Chang. "Temporal action localization in untrimmed videos via multi-stage cnns." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[12] Yeung, Serena, et al. "End-to-end learning of action detection from frame glimpses in videos." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[13] Wang, Limin, et al. "Temporal segment networks: towards good practices for deep action recognition." *European Conference on Computer Vision*. Springer International Publishing, 2016.
[14] Mansimov, Elman, Nitish Srivastava, and Ruslan Salakhutdinov. "Initialization strategies of spatio-temporal convolutional neural networks." *arXiv preprint arXiv:1503.07274* (2015).
[15] Molchanov, Pavlo, et al. "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[16] Cao, Congqi, et al. "Action Recognition with Joints-Pooled 3D Deep Convolutional Descriptors."
[17] M. Ibrahim,et al. A hierarchical deep temporal model for group activity recognition. *arXiv preprint arXiv:1511.06040*, 2015 .
[18] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in neural information processing systems*. 2014.