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

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, University of Barcelona, Barcelona, Spain | [3]Computer 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

Abstract. In this paper...

We present:

- A **survey** on current deep learning methodologies (action/gesture recognition)
- A **taxonomy** summarizing 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. Questions that remain opened:

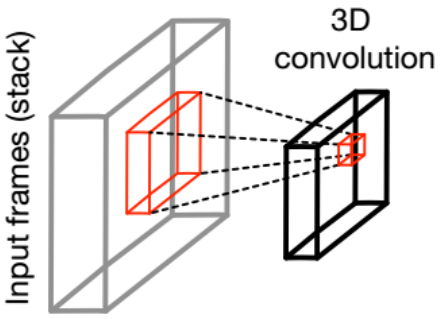
- How to deal with **temporal information**. We investigated works that go beyond averaging class score predictions on individual frames for video prediction
- How to train deep models on **small datasets**
- Whether they are used in **combination with hand-crafted features**
- Which are the most successful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

Taxonomy

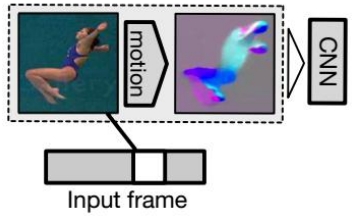
Architectures

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

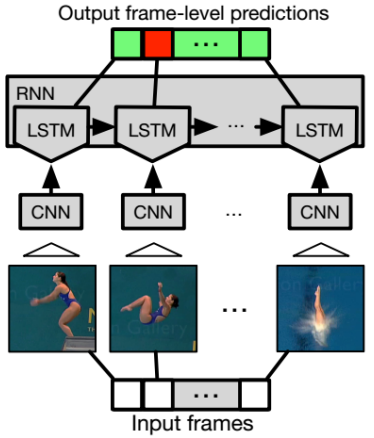
- ❑ **3D convolutions**
- ❑ **Motion-based approaches**
- ❑ **Sequential models**



3D convolutions



Motion-based approaches



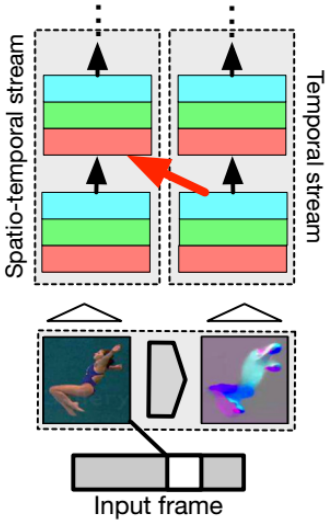
Sequential models

Fusion strategies

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, late fusion, slow fusion**



Slow fusion

State-of-the-art methods and results

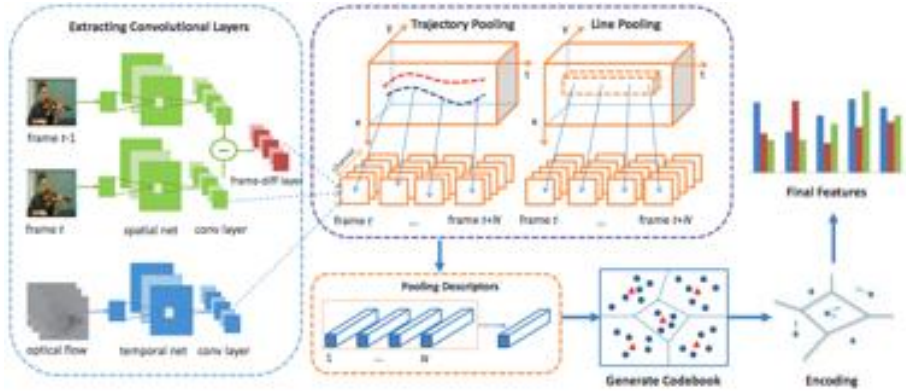
UCF-101 results

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%

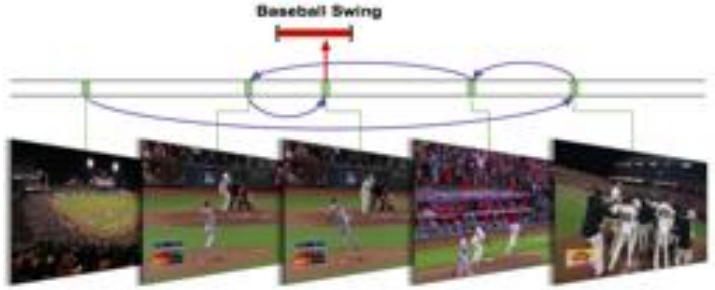
THUMOS'14 results

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

[2]



[12]



Discussion

We bring some discussion: on which are the better **temporal modeling** strategies, on problems that arise training with **small datasets**, on the exploitation of **hand-crafted features** in hybrid approaches, and on **future trends and research directions**