Late Fusion of Bayesian and Convolutional models for Action Recognition

Camille Maurice[1], Francisco Madrigal[1] et Frédéric Lerasle[12]
[1]LAAS-CNRS, [2]University Paul Sabatier, Toulouse - contact: cmaurice@laas.fr

Problem Definition and Contributions

- **Goal**
  - To **recognize actions** on videos where a person performs several sequential actions with objects.
- **Our Contributions**
  - Addition of **recurrent layer** to a 3D CNN to take into account action transitions
  - A **hybrid** approach based on **late fusion**

Key Statements

- A Bayesian approach with great results on **specific** actions
- A **possible synergy** between a deep-learning and Bayesian approaches
- **Temporal consistency** within an action and throughout the sequence
- Action sequences importance to **leverage ambiguities** thus improve performances

Problem Definition and Contributions

- Goal
  - To **recognize actions** on videos where a person performs several sequential actions with objects.
- Our Contributions
  - Addition of **recurrent layer** to a 3D CNN to take into account action transitions
  - A **hybrid** approach based on **late fusion**

Key Statements

To reach

70% → To move

18% → To open

12% → To close

~~0%~~ To eat

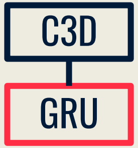
Models Characteristics

- **Bayesian Model (ANBM) [1]**
  ANBM
  - Models skeletons, skeletons-objects and object-object interactions in a joint probability
  - Models action transitions
- **C3D [5]**
  C3D
  - 3D convolutions to learn spatio-temporal features on video clips
  - Does not model action transitions

[1] A new Bayesian Model for Action Recognition, C. Maurice *et al.*
[5] Learning spatiotemporal features with 3d convolutional networks, D. Tran *et al.*

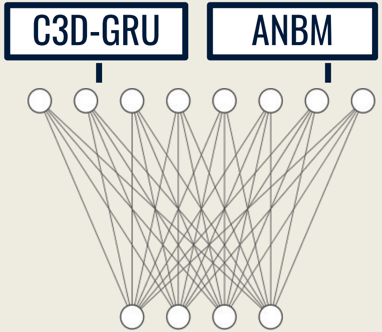
C3D-GRU

- **C3D-GRU**
  - Recurrent layer with **memory** cell to capture time dependencies
  - Gated Recurrent Unit, a recurrent layer suitable for **short sequences**
  - Freeze C3D weights, then train the GRU layer with **2 successive** clips as input



Late Fusion with a Dense Layer

- **Predictions concatenation**
- Connection to a **fully-connected** layer with **soft-max** activation
- Enable a **late fusion** with correlations between predictions



Results

Dataset	Approaches	Accuracy
CAD-120	GEPHAPP[2]	79.4
	ANBM[1]	82.2
	GPNN[3]	87.3
	Ours	86.1
Watch-n-Patch	PoT[4]	49.9
	ANBM[1]	76.4
	GEPHAPP[2]	84.8
	Ours	93.0

- **2 public datasets**

Results

	0	1	2	3	4	5	6	7	8	9
0	0.89	0.01	0.01	0.07	0.00	0.02	0.00	0.00	0.00	0.00
1	0.00	0.89	0.00	0.00	0.03	0.05	0.00	0.00	0.02	0.01
2	0.00	0.05	0.87	0.00	0.01	0.04	0.00	0.00	0.02	0.01
3	0.00	0.00	0.00	0.77	0.00	0.00	0.23	0.00	0.00	0.00
4	0.05	0.02	0.01	0.00	0.54	0.38	0.00	0.00	0.00	0.00
5	0.08	0.01	0.01	0.00	0.38	0.52	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.25	0.00	0.00	0.75	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.06	0.04
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.81	0.01
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.04	0.82

ANBM

	0	1	2	3	4	5	6	7	8	9
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.98	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00
2	0.00	0.01	0.97	0.00	0.00	0.02	0.00	0.00	0.01	0.00
3	0.00	0.00	0.03	0.81	0.06	0.10	0.00	0.00	0.00	0.00
4	0.00	0.00	0.09	0.00	0.91	0.00	0.00	0.00	0.00	0.00
5	0.00	0.03	0.01	0.00	0.00	0.89	0.04	0.01	0.00	0.02
6	0.00	0.02	0.05	0.05	0.00	0.07	0.80	0.02	0.00	0.00
7	0.04	0.02	0.00	0.02	0.00	0.05	0.00	0.86	0.00	0.02
8	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.09	0.78	0.06
9	0.00	0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.12	0.75

C3D-GRU

	0	1	2	3	4	5	6	7	8	9
0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.97	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01
2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.94	0.00	0.00	0.06	0.00	0.00	0.00
4	0.00	0.00	0.09	0.00	0.91	0.00	0.00	0.00	0.00	0.00
5	0.01	0.03	0.01	0.00	0.01	0.94	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.09	0.00	0.00	0.91	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.05	0.05
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.81	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.81

FUSION

Conclusion and Future Works

- Performance gain, particularly in **under-represented** classes
- Performance gain when the sources of error are different

https://www.youtube.com/watch?v=7txCiHx3OwA

Acknowledgements

- This work has been partially supported by Bpifrance within the French Project LinTO and funded by the French government under the Investments for the Future Program (PIA3).