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

70% → To move
18% → To open
12% → To close
0% → To eat

To reach

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** — 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

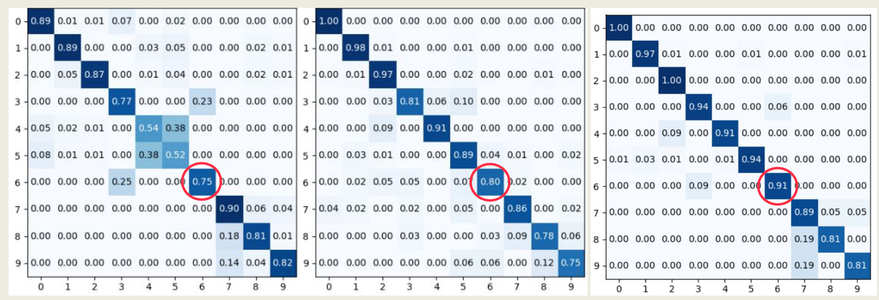
- **Predictions concatenation**
- Connection to a **fully-connected** layer with **soft-max** activation
- Enable a **late fusion** with correlations between predictions

Results

- Evaluation on two **public** datasets : CAD-120 and Watch-n-Patch

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

- Confusion matrices on Watch-n-Patch



ANBM C3D-GRU Fusion

[2] A generalized earley parser for human activity parsing and prediction, S.Qi *et al.*
[3] Learning human-object Interactions by graph parsing neural networks, S.Qi *et al.*
[4] Pooled motion features for first-person videos, M. S. Ryoo *et al.*

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).