2D Deep Video Capsule Network with Temporal Shift for Action Recognition

T. Voillemin - H. Wannous - J-P. Vandeborre

Université de Lille

IMT Lille Douai
École Mines-Télécom
IMT-Université de Lille

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

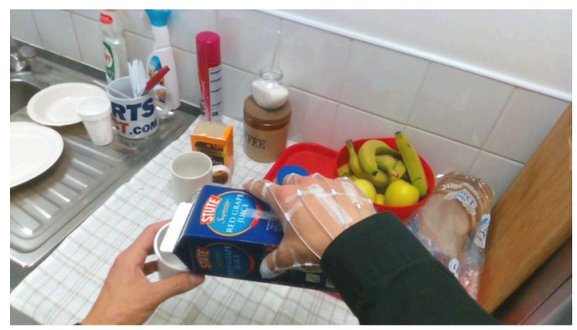
ICPR 2020

10-15 January 2021

Introduction

Motivations & challenges

- Action Recognition for Virtual/Augmented reality devices

Garcia-Hernando et al., *CVPR*, 2018


De Smedt et al., *SHREC*, 2017

- Light but efficient algorithm for real-time applications
- Working with easily accessible data as RGB and depth video stream

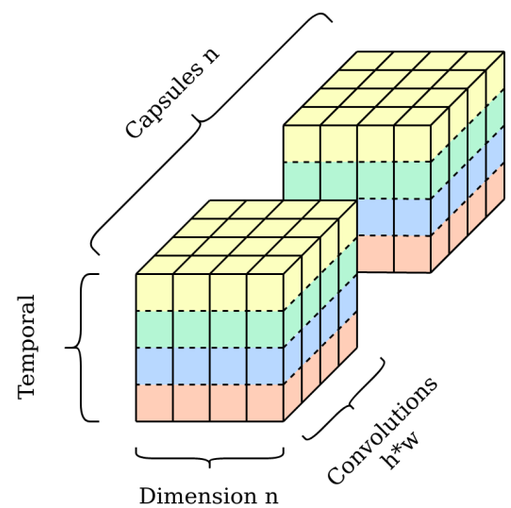
Introduction

Existing approaches

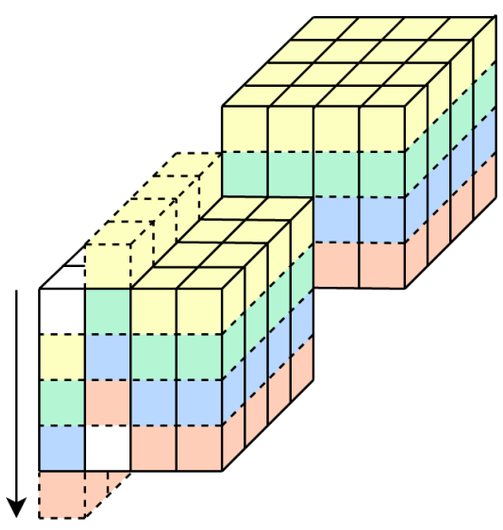
- For lightness :
  - Capsule Network (Sabour et al., *ANIPS*, 2017)
- For efficiency :
  - DeepCaps (Rajasegaran et al., *CVPR*, 2019)
  - Temporal Shift Module (Lin et al. *ICCV*, 2019)

Proposed Method

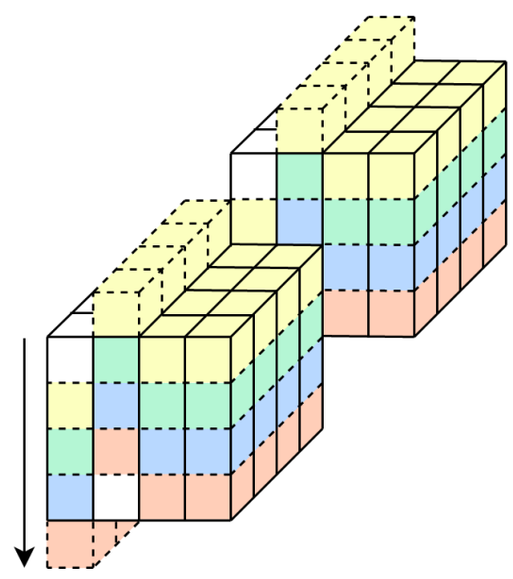
Temporal Shift on Capsule Layer



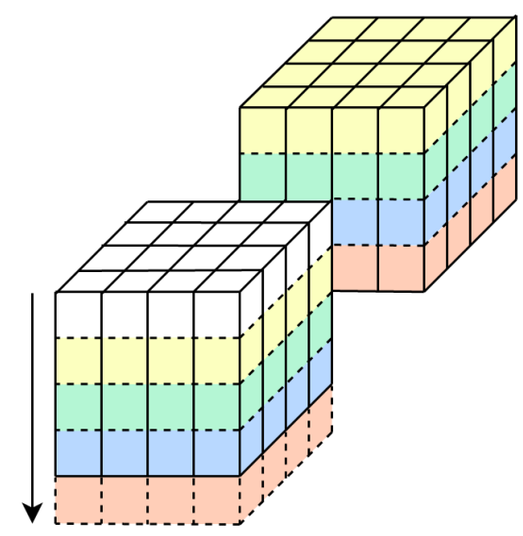
Original Capsule Layer



Shift of first kernels on first capsule



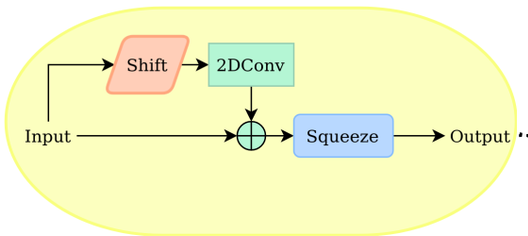
Shift of first kernels on every capsules



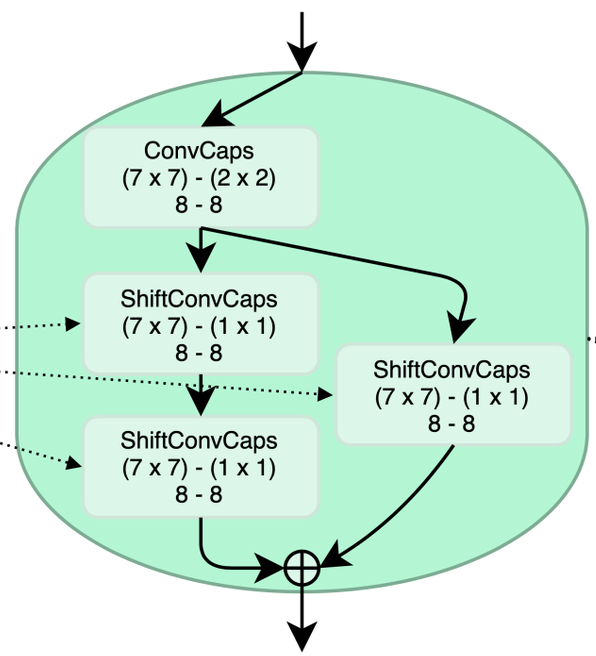
Shift of every kernels on first capsule

Proposed Method

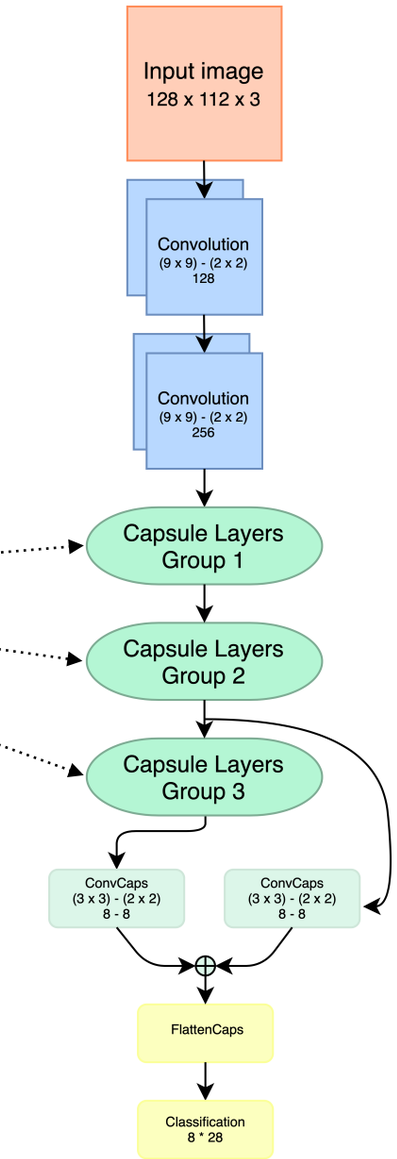
Network Architecture



ShiftConvCaps



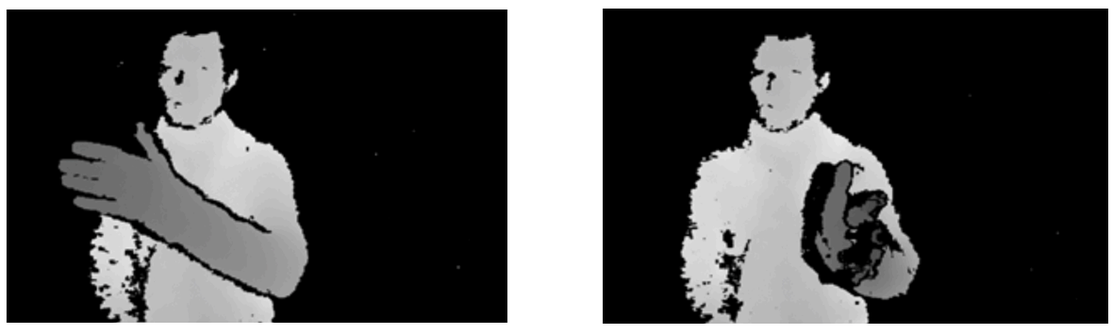
Capsule Layer Group



Full Network

Experiments

Accuracy comparaison over DHG dataset



De Smedt et al., *SHREC*, 2017

Method	Parameters	Accuracy (%)
TSM [2]	24 M	58.66
2D 3DCNN Fusion [3]	140 M	**74.41**
Ours	**7 M**	**68.98**

Comparaison on DHG28 [5]

[1] Feichtenhofer et al., *CVPR*, 2016
[2] Lin et al., *ICCV*, 2019
[3] Zhang et al., *Electronics*, 2019
[4] Garcia-Hernando et al., *CVPR*, 2018
[5] De Smedt et al., *SHREC*, 2017

Experiments

Accuracy comparaison over FPHA dataset

Method	Parameters	Accuracy (%)
Two stream-color [1]	46 M	61.56
Two stream-flow [1]	46 M	69.91
Two stream-all [1]	181 M	75.30
TSM [2]	24 M	71.57
Ours	**4 M**	**76.72**

Comparaison on FPHA [4]

Garcia-Hernando et al., *CVPR*, 2018

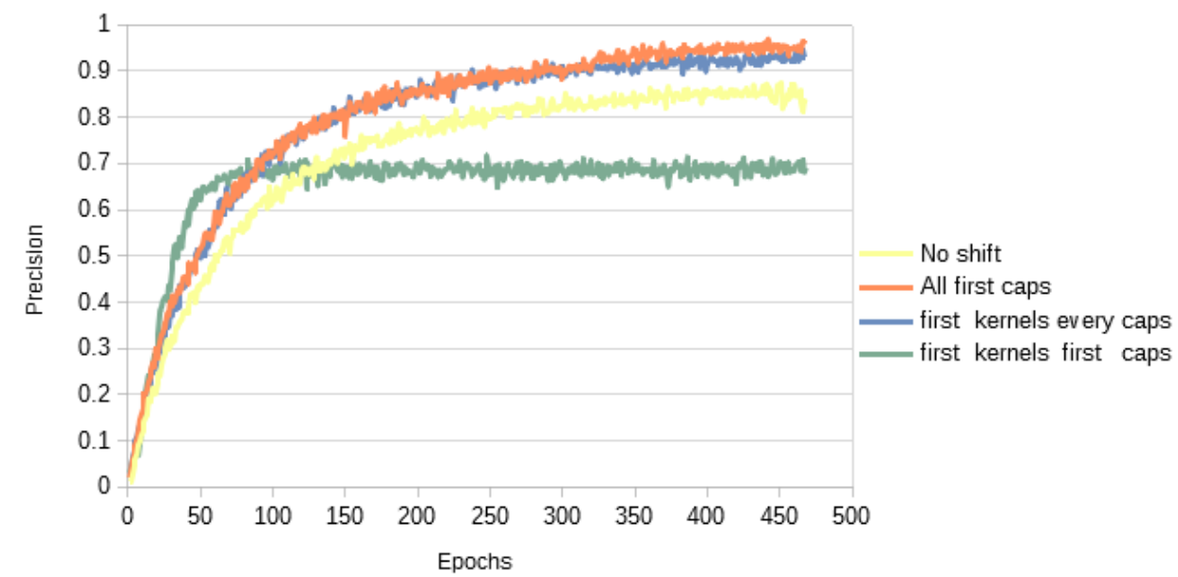
[1] Feichtenhofer et al., *CVPR*, 2016
[2] Lin et al., *ICCV*, 2019
[3] Zhang et al., *Electronics*, 2019
[4] Garcia-Hernando et al., *CVPR*, 2018
[5] De Smedt et al., *SHREC*, 2017

Experiments

Accuracy comparaison over FPHA dataset

Temporal Shift	Accuracy (%)
No shift	70.01
Shift first kernels on first capusles	64.14
Shift first kernels on every capsules	**74.33**
Shift every kernels of first capsules	**76.72**

Final testing accuracy



Training accuracy evolution over epochs

Conclusion

Contributions & perspectives

- First 2D Capsule Network for video understanding
- Implementation of temporal shift over capsule layer
- Outperform or near state-of-the-art with 10 to 40 times less parameters

- Adapt caspules for hand skeleton data
- Use it in real-time directly on augmented/virtual reality devices

Thank you for your attention

T. Voillemin - H. Wannous - J-P. Vandeborre

Université de Lille

IMT Lille Douai
École Mines-Télécom
IMT-Université de Lille

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille