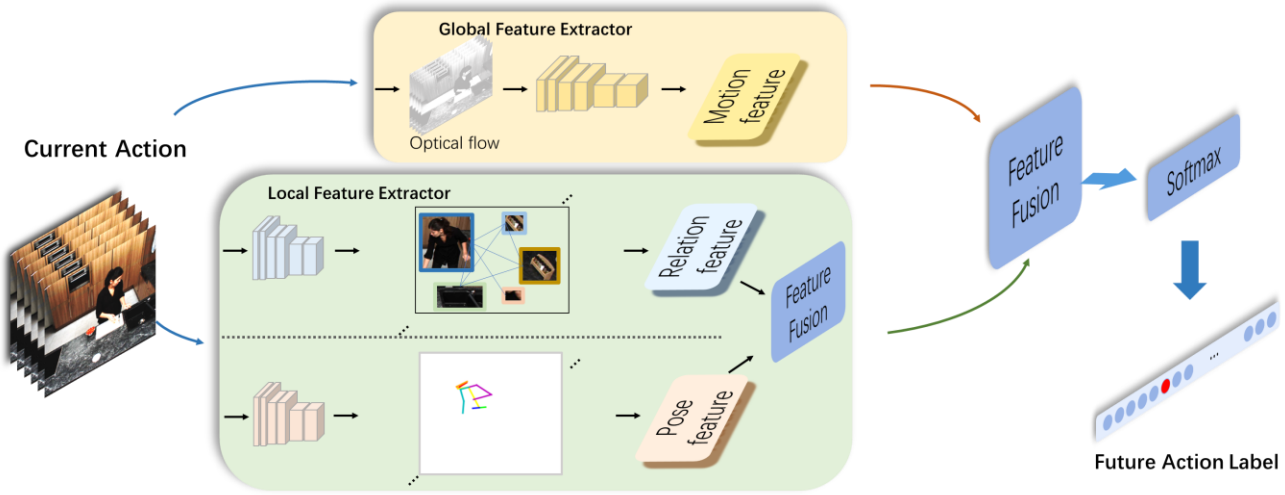
Weakly Supervised Action Anticipation without Object Annotations

Yi ZHONG, Jia-Hui PAN, Haoxin LI, Wei-Shi ZHENG

Frontiers of Computer Science, DOI: 10.1007/s11704-022-1167-9

Problems & Ideas

- Problems of methods of predicting next different action:
  - Those models are under strong supervised learning setting, integrating context labels directly, such as the tools, ingredients and containers observed in the action.
  - Fine-grained labelling requires considerable effort in a large-scale video repository.
- Ideas: A weakly supervised method integrates global motion and local fine-grained features to predict next action label without the need for specific scene context labels.



The model is divided into two types of part models for different capacities. The yellow part is Global Motion branch. The upper of the green part is Appearance Relation branch and the other one is Human Skeleton branch.

Main Contributions

- Contributions:
  - A weakly supervised learning approach to obtain abundant information for action anticipation task, which can reduce a lot of elaborated labelling efforts;
  - A fusion framework with global motion features and local fine-grained features, including appearance relations and human pose.

Method	Top-1 Accuracy (%)	Top-5 Accuracy (%)
Contexts [1]	33.1	-
CNFAS [2]	33.7	-
Contexts (I3D)	34.82	59.11
Ours	**39.67**	**66.12**

Results on the MPII-Cooking dataset

Method	Top-1 Accuracy		Top-5 Accuracy	
	VERB	ACTION	VERB	ACTION
Contexts (I3D)	27.41	05.26	73.01	15.22
I3D	27.82	05.62	74.22	16.89
GAT	29.29	07.46	74.88	20.67
Ours	**31.13**	**08.41**	**75.82**	**22.34**

Results on the EPIC-Kitchens dataset.