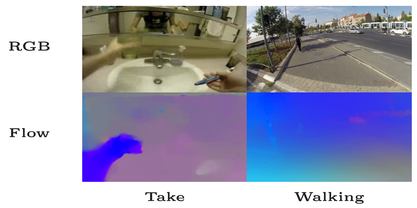
Making Third Person Techniques Recognize First-Person Actions in Egocentric Videos

Sagar Verma, Pravin Nagar, Divam Gupta, and Chetan Arora

sagar.verma@centralesupelec.fr, pravinn@iiitd.ac.in, divam14038@iiitd.ac.in, chetan@iitd.ac.in

Problem Statement

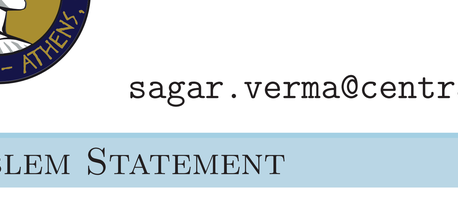
DNN trained on third-person actions do not adapt to egocentric actions due to a large difference in size of visible objects. Another complexity is multiple action categories. This work unifies the feature learning for multiple action categories using a generic two-stream architecture.



Actions with hand-object interaction(take) and without(walking) in two different view streams.

Contributions

1. Deep neural network trained on third person videos do not adapt to egocentric action due to large difference in size of the visible objects.



After cropping and resizing the objects become comparable to the objects in third person videos.

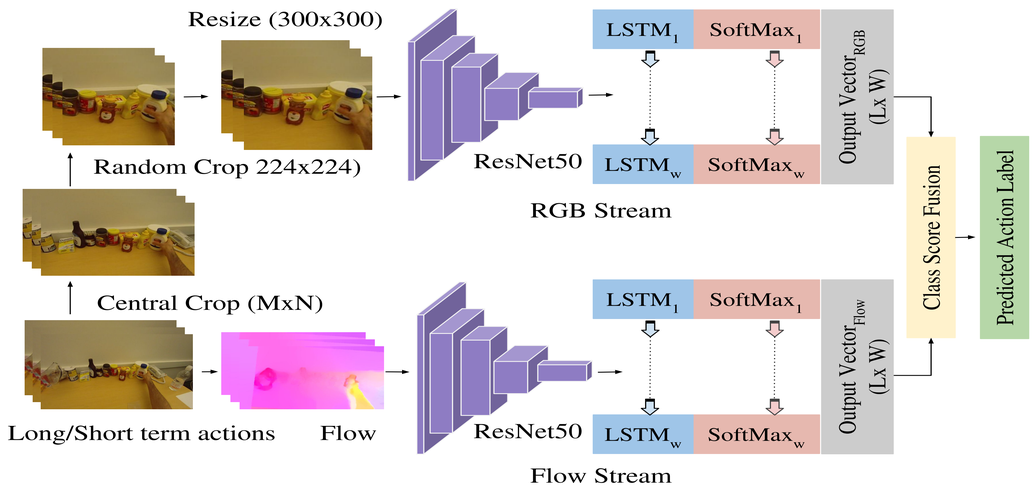
2. We propose curriculum learning by merging similar but opposite actions while training CNN.

3. Proposed framework is generic to all categories of egocentric actions.

Related Work

Earlier works on first-person action recognition use hands and objects as important cues.[1, 2] On the other end many works only use motion information for first-person action recognition.[3, 4] State of the art (SoTA) techniques focus only on one specific category of action classes.

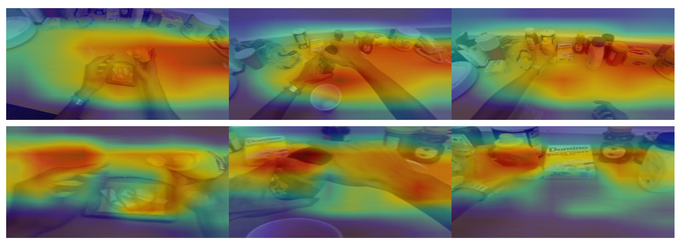
Proposed Architecture



Results and Discussion

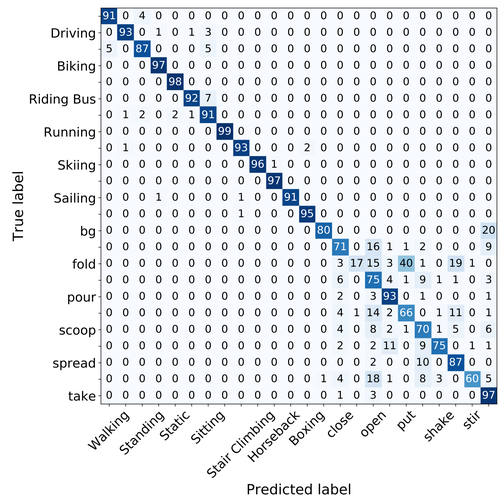
Dataset	Subjects	Frames	Classes	Accuracy Current	Accuracy Ours
GTEA [1]	4	31,253	11	68.50[5]	82.71
EGTEA+ [1]	32	1,055,937	19	NA	66
Kitchen [6]	7	48,117	29	66.23[5]	71.92
ADL [2]	5	93,293	21	37.58[5]	44.13
UTE [7]	2	208,230	21	60.17[5]	65.12
HUJI [8]	NA	1,338,606	14	86[8]	93.92

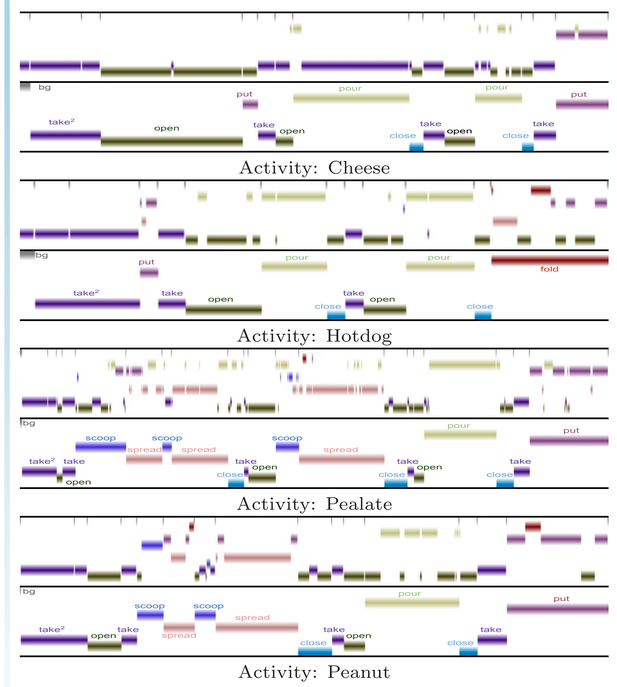
Accuracy comparison of our method with SoTA and statistics of egocentric video datasets



Top and bottom rows show the visualization of normal and resized inputs respectively for 'close', 'open', and 'take' actions column-wise.

Applicability in real life setting where different action categories are present:
To validate the applicability of our method, we use mixed samples from GTEA [1] and HUJI [8] dataset. From the confusion matrix it is evident that the proposed network does not seem to have any confusion in the different ca8tegory of actions.





Top and bottom of each subfigure shows predicted and ground truth sequence respectively.

References

[1] Alireza Fathi, Xiaofeng Ren, and James M Rehg, "Learning to recognize objects in egocentric activities," in *CVPR*, 2011.

[2] Hamed Pirsiavash and Deva Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012.

[3] Kris Makoto Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *CVPR*, 2011, pp. 3241–3248.

[4] Suriya Singh, Chetan Arora, and C. V. Jawahar, "Trajectory aligned features for first person action recognition," *Pattern Recognition*, vol. 62, pp. 45–55, 2016.

[5] Suriya Singh, Chetan Arora, and C V, Jawahar, "First person action recognition using deep learned descriptors," in *CVPR*, 2016, pp. 2620–2628.

[6] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert, "Temporal segmentation and activity classification from first-person sensing," in *CVPRW*, 2009, pp. 17–24.

[7] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman, "Discovering important people and objects for egocentric video summarization.," in *CVPR*, 2012, pp. 1346–1353.

[8] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora, "Compact cnn for indexing egocentric videos," in *WACV*. IEEE, 2016, pp. 1–9.

Acknowledgement

This work has been supported by Infosys Center for Artificial Intelligence, Visvesaraya Young Faculty Research Fellowship, and Visvesaraya Ph.D. Fellowship from Government of India. We thank Inria and CVN Lab at Centralesupelec to support travel for Sagar Verma.