What do 15,000 object categories tell us about classifying and localizing actions?

Mihir Jain[†], Jan C. van Gemert[†], Cees G. M. Snoek[†⋆]
[†]University of Amsterdam. [⋆]Qualcomm Research Netherlands.

In this paper we ask ourselves the question: *"What do 15,000 object categories tell us about classifying and localizing actions?"* Whereas motion is the key ingredient in modern approaches, we assess the benefits of having objects in the video representation. Rather than considering a handful of carefully selected and localized objects, we conduct an empirical study on the benefit of encoding 15,000 object categories for action using 6 datasets totaling more than 200 hours of video and covering 180 action classes. Our study is inspired by the example set by Deng *et al.*[1] for image categorization. Our key contributions are *i)* the first in-depth study of encoding objects for actions, *ii)* we show that objects matter for actions, and are often semantically relevant as well. *iii)* We establish that actions have object preferences. Rather than using all objects, selection is advantageous for action recognition. *iv)* We reveal that object-action relations are generic, which allows to transfer these relationships from the one domain to the other. And, *v)* objects, when combined with motion, improve the state-of-the-art for both action classification and localization.

Object and motion representation. As object representation, we compute the likelihood of the presence of the 15k object categories in each frame of the considered videos. We use an in-house implementation of a Krizhevsky style cuda-convnet with dropout [4]. The final ∼15k is obtained by averaging the objects responses over frames. Robust motion descriptors along the dense trajectories encoded with Fisher vector is used as the motion representation.

Objects matter for actions: We show that objects matter for actions with the help of visualizations and quantitative experiments. For visualization, we use tag-clouds and a heat-map of objects responses for action classes. We find that the most responsive objects are often semantically relevant as well. In our classification experiments, we find the object representation to be complimentary to modern motion encodings and more so when the actions interact with objects. Another interesting observation is that the object responses in the proximity of actions are most informative.

Actions have object preference: For a given dataset of n action classes, we assign a set of the top R most responsive object categories, $top_R(c_j)$, to each action class j. The union of these n sets of object categories, gives us a set of preferred objects for the given set of action classes, $\Gamma(R)$. In Figure 1, we evaluate the impact of object preference on action classification by varying the value of R in light of a representation consisting of (a) objects only and (b) objects with motion on the THUMOS14 validation set. The accuracies are computed for different values of R, *i.e.* starting with no objects, then progressively adding the most responsive object categories for the given dataset, till all the object categories are used.

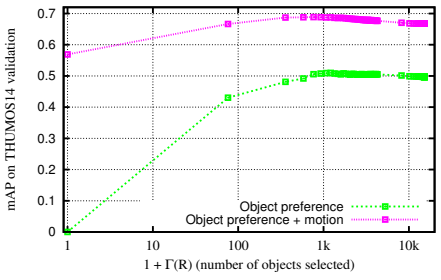

Figure 1: Actions have object preference, selecting characteristic objects for a given set of actions is crucial.

Object-action relations are generic: We evaluate if this knowledge of characteristic objects learned from one dataset can be transferred to another dataset. For this we conduct experiments on HMDB51 and UCF101 as they have 12 action classes in common. We learn the preferred set of objects from the training sets of HMDB51 and another preferred set from UCF101. Then we use these sets for the object representation of videos in the test set of HMDB51. We compare the impact of the representations for these two transfers in Table 1. Interestingly, learning the characteristic set of objects on UCF101, leads to a better mean accuracy on HMDB51 which implies that object-action relations are generic.

Method	Motion	+ Object categories selected from	
		HMDB51	UCF101
Mean accuracy on HMDB51	83.6%	87.5%	88.1%

Table 1: The characteristic object categories learned from the training sets of HMDB51/UCF101 transfer to the test set of HMDB51.

Objects improve the state-of-the-art: Adding object encodings improves the state-of-the-art in both action classification and localization. In Table 2, we compare our results with the best reported results till date on four challenging datasets. Results for localization are shown in Figure 2. We conduct this experiment on UCF Sports using the tubelet action proposals of Jain *et al.* [2].

Method	UCF101	THUMOS14	Hollywood2	HMDB51
Best reported	87.7%	71.0%*	**73.7%**	66.8%
Ours	**88.5%**	**71.6%**	66.4%	**71.3%**

Table 2: Objects improve the action classification state-of-the-art. *Our winning approach at THUMOS14 [3].

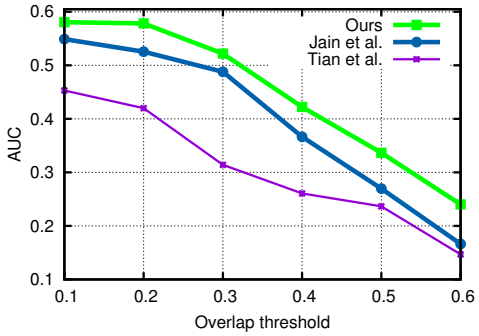

Figure 2: Objects outperform the action localization state-of-the-art methods, Tubelets [2] and SDPM [5].

Acknowledgments This research is supported by the STW STORY project, the Dutch national program COMMIT, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

[1] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[2] M. Jain, J. C. van Gemert, H. Jégou, P. Bouthemy, and C. G. M. Snoek. Action localization with tubelets from motion. In *CVPR*, 2014.

[3] M. Jain, J. C. van Gemert, and C. G. M. Snoek. University of Amsterdam at THUMOS Challenge 2014. In *ECCV workshop*, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[5] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.