Harnessing Object and Scene Semantics for Large-Scale Video Understanding Supplemental Material

Zuxuan Wu[†], Yanwei Fu[§], Yu-Gang Jiang[†], Leonid Sigal[§]
[†]Shanghai Key Lab of Intel. Info. Processing, School of Computer Science, Fudan University
[§]Disney Research
{zxwu, ygj}@fudan.edu.cn, {yanwei.fu, lsigal}@disneyresearch.com

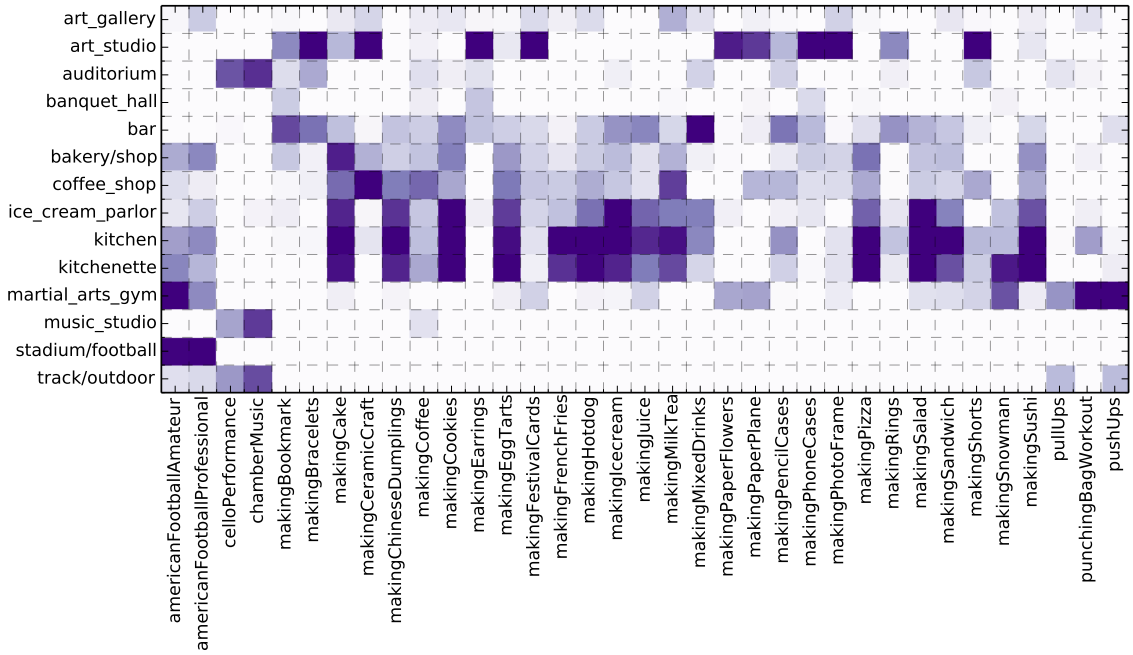


Figure 1: The visualization of a part from the Π^S learned on FCVID, where each entry denotes the response score between each scene and video class pair.

1. Interpretability of OSR

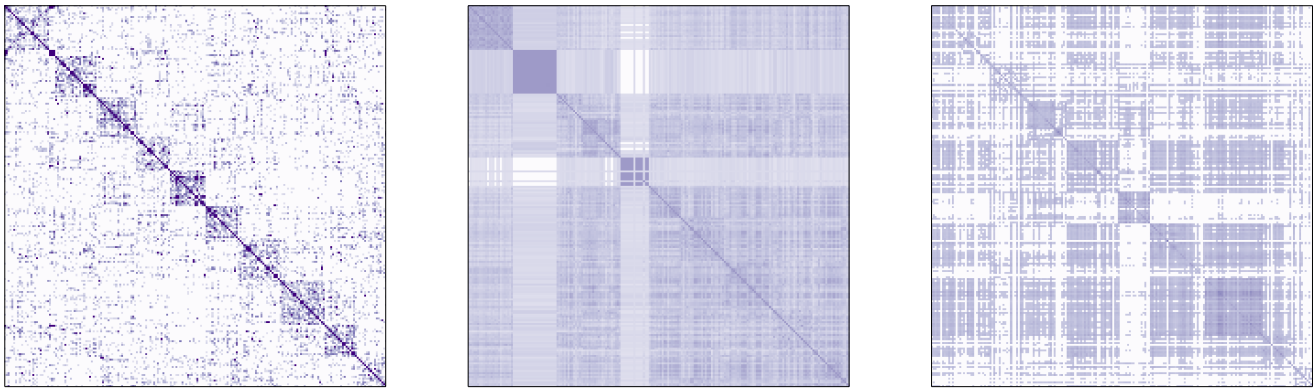
To demonstrate the derived relationship between video classes and scenes, we visualize part of Π^S from FCVID in Figure 1, where each entry indicates the likelihood of the video categories (activities) occurring under the corresponding scene. As can be seen from the figure, most categories related to food (*e.g.*, "making cake" and "making French fries") have high response scores from "kitchens" and "kitchenette", while some DIY classes most likely to take place in "art studios". It is also interesting to see that the derived relationship can effectively differentiate between indoor (*e.g.*, "push ups") and outdoor sports (*e.g.*, "American football").

2. Comparisons with

	Mihir *et al.*	Ours
Objecs	65.6	69.4
Overall	88.1	86.4

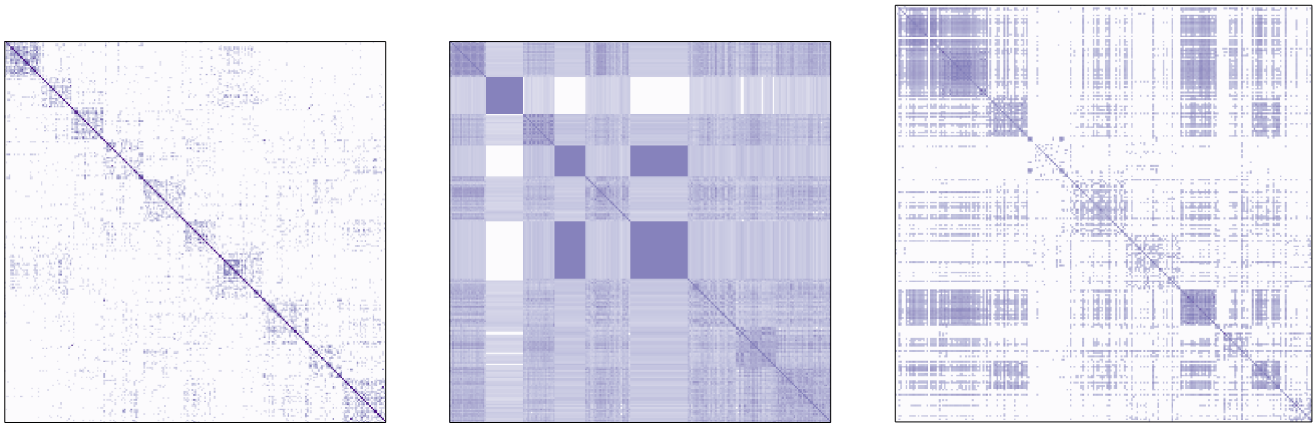
Table 1: Comparisons with Mihir

We also compare our results with Mihir *et al.* [1] on UCF101 [3]. Our object stream achieves 4% performance gain, while the overall performance is 1.7% lower compared to theirs. Note that the results are not directly comparable, since different CNN models are utilized and their approach relies on state-of-the-art handcrafted features to capture mo-



(a) Class similarity generated by OSR (b) Class similarity generated by word vectors. (c) Class similarity based on ground-truth.

Figure 2: Class similarity generated by different methods on ActivityNet.



(a) Class similarity generated by OSR (b) Class similarity generated by word vectors. (c) Class similarity based on ground-truth.

Figure 3: Class similarity generated by different methods on FCVID.

tion information. We would like to underline that our framework is designed to harness high-level objects/scene semantics with a generic feature stream to account for low-level information, which can be easily replaced with more advanced networks such as Two-Stream CNN [2], and hence our framework is more flexible.

3. The Effectiveness of OSR for Class Similarity Discovery

Intuitively, related classes should possess similar OSR representations, since they usually contain the same set of objects and occur under similar scenes. Therefore, we compute the similarity matrix of all classes to validate the derived OSR and compare with that generated by word vectors and ground truth. Figure 3 and Figure 2 demonstrate the results on ActivityNet and FCVID respectively (Normalized Cut is performed for better visualization). The figures clearly demonstrate that the OSR can effectively discover class structures on both datasets. Comparing OSR with word vectors on both datasets, we can see that visually learned relationships prove to be superior than obtained on text corpus.

References

[1] M. Jain, J. C. van Gemert, and C. G. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, 2015. 2

[2] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 2

[3] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012. 2