Abstract

Scene Classification has been an active area of research in Computer Vision. The goal of scene classification is to classify an unseen image into one of the scene categories, *e.g*. beach, cityscape, auditorium, etc. Indoor scene classification in particular is a challenging problem because of the large variations in the viewpoint and high clutter in the scenes. The examples of indoor scene categories are corridor, airport, kitchen, etc. The standard classification models generally do not work well for indoor scene categories. The main difficulty is that while some indoor scenes (*e.g*. corridors) can be well characterized by global spatial properties, others (*e.g*. bookstores) are better characterized by the objects they contain. The problem requires a model that can use a combination of both the local and global information in the images.

Motivated by the recent success of the Bag of Words model, we apply the model specifically for the problem of Indoor Scene Classification. Our well-designed Bag of Words pipeline achieves the state-of-the-art results on the MIT 67 indoor scene dataset, beating all the previous results. Our Bag of Words model uses the best options for every step of the pipeline. We also look at a new method for partitioning of images into spatial cells, which can be used as an extension to the standard Spatial Pyramid Technique (SPM). The new partitioning is designed for scene classification tasks, where a non-uniform partitioning based on the different regions is more useful than the uniform partitioning.

We also propose a new image representation which takes into account the discriminative parts from the scenes, and represents an image using these parts. The new representation, called Bag of Parts can discover parts automatically and with very little supervision. We show that the Bag of Parts representation is able to capture the discriminative parts/objects from the scenes, and achieves good classification results on the MIT 67 indoor scene dataset. Apart from getting good classification results, these blocks correspond to semantically meaningful parts/objects. This mid-level representation is more understandable compared to the other low-level representations (*e.g*. SIFT) and can be used for various other Computer Vision tasks too.

Finally, we show that the Bag of Parts representation is complementary to the Bag of Words representation and combining the two gives an additional boost to the classification performance. The combined representation establishes a new state-of-the-art benchmark on the MIT 67 indoor scene dataset. Our results outperform the previous state-of-the-art results by 14%, from 49.40% to 63.10%.