**Paper Title:** Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories

**Paper Summary:**

This paper can be regarded as one of the most important papers related to image / scene classification based on the “Bag of Words” (BoW) scheme. It inspired a lot of subsequent literatures in an effort to improve on top of this method.

In this paper the author described and discussed a very simple yet effective image/scene classification method - called spatial pyramid matching - which basically computes the BoW matching score in different levels of coarseness with regularly divided grids on an image plane. The final matching score between two images is defined as a weighted average of all the matching scores computed at each level and each grid within the same level.

The intuition is that BoW focuses too much on order-less matching and the proposed method tries to find a trade-off between the BoW and the exact matching. If spatially finer matches are found between two images then certainly this should give a bonus to the matching level of the two images. In BoW however, this is ignored.

Despite the unbelievable simplicity, the proposed method achieved the state of the art performance on the Fifteen Scene Categories and Caltech 101 at that time. It also achieved reasonable performance on the Graz Dataset. The paper also gave very detailed and intuitive discussions on why the proposed method works and why not on certain dataset / images.

**Strength:**

1. The most salient feature of this paper is that its proposed method is so simple that it can be easily reproduced by many others to be verified. This is exactly the kind of “good paper and method” I personally like: simple, intuitive yet highly effective.
2. The paper is technically correct and the conducted experiments clearly demonstrated the effectiveness. A significant gap between the performance of the method and the previous state of the art can be observed.
3. The paper is highly inspiring because shows us the huge potential of performance increase even with very simple strategies to handle the order-less level and incorporate spatial information.
4. The paper gives very clear explanation and discussion why the proposed method works or not on certain examples.

**Weak Points:**

**1) Images vs. Objects**

This paper provides a good intuition and direction we need to look into: **taking certain level of structural information into account helps**. But the proposed way through which the structural information is organized is in this paper is clearly too naïve and rigid. In other words, the proposed scheme cannot cope with many common cases where global structural information is indeed much orderless, as is illustrated bellow in Fig. 1.



Fig. 1. An example showing orderless scenes in the real world.

We can generalize the above situation in a simple toy example. Consider the following example in Fig. 2 where we have two images containing exactly the same sub-windows except their orders are different. In this case, spatial pyramid matching clearly works not so well since the order of the two histograms to be matched are also different from each other under spatial pyramid matching. The authors also noticed this point and mentioned in their paper.



Fig. 2. A toy example showing the failure situation of spatial pyramid matching.

The essence of the weakness lies in the fact that by nature scene information is not only embedded in image level structures, but is also embedded in object levels. This brings us to the philosophy of “Images verses Objects”. Some scenes are more globally structured and bias more towards image level features. Here are two sets of examples:

 

Street Views

 

Beautiful Seas

You see all of them are nicely and globally structured. But for some other scenes, object seems to be playing a more important role in predicting the scene labels. See the following examples:

 

Tennis Games

The most reliable cues to classify this scene are tennis rackets and balls. In addition face recognition sometimes may also help to figure out the scene context. All these are object level evidences. And unfortunately, objects do not always stay at the same positions.

Which ultimately is the most important one that defines a scene? In my personal opinion, I tend to choose objects over images. This is because image-level structural information sometimes is simply too difficult to generalize, while generalizing objects can be relatively easier. In addition, recognizing objects has the potential to be composed into image level structures.

Given the above discussion, a more reasonable formulation for object-oriented or cluttered scenes might be performing spatial pyramid matching on object-level patches and match the images in a less rigid way.

**2) Image Partitioning**

The proposed way of partitioning an image clearly is not invariant to scale, translation and rotation. Recently there have been papers proposing “Spatial-Bag-of-Features” which encode geometric information and is invariant to scale, translation and rotation. They introduced two different ways for image partition. The first one is linear ordered bag-of-features, in which image is partitioned into straps along a line with an arbitrary angle. The second one is circular ordered bag-of-features, in which a center point is given and then the image is evenly divided into several sectors with the same radian. By enumerating different line angles (ranging from 0◦ to 360◦) and center locations, a family of linear and circular ordered bag-of-features can be obtained. See the paper “Spatial-bag-of-features” by Cao et al. in CVPR 2010 for more details.

**3) Dataset Biases**

From the paper we also know that the paper has a certain taste of datasets. While these datasets consists considerable amount of images at that time, nowadays we know they are in some sense biased. For example the fifteen scene categories typically consist of scenes with nice viewing angles and global structures, which clearly favor spatial pyramid matching. The bias may result from the relatively restricted locations (MIT), the (fixed) way they select images, as well as the (fixed) way a photographer takes photos.

Caltech 101 shows the same problem too. The dataset seldom contains images with cluttered backgrounds common encountered in real life.

**Inspirations:**

There are many inspirations aroused by this paper. The first and most important one is: Always optimize and make trade-offs in your problems. Trade-offs often can bring you performance gain.

Second, despite the fact there are many weak points of this method, it is not to say it will become useless ultimately, or only represents a “wrong way which happened to work well on specific scenarios”. My personal view is that matching tasks for human should not be a single, flat process. Rather, it is a hierarchical one with multiple stages. Humans certainly are familiar with many canonical scenes which require virtually no effort to interpret. Some of the processes are even finished subconsciously without deliberate understanding. These are the typical scenarios where simple methods work the best, such as nearest neighbors and spatial pyramid matching. Other scenarios require more complicated understanding processes and models that generalize better. For example, unfamiliar scenes where people need to distinguish objects first in order to figure out the overall context. Both are indispensable to formulate a good vision system.

**Conclusion:**

Our world is complex. This is definitely not a paper to present a final solution, yet it is a good paper which marks important efforts made by the vision people. As Winston Churchill said: “Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.”