# BAG-OF-FEATURES REPRESENTATIONS USING SPATIAL VISUAL VOCABULARIES FOR OBJECT CLASSIFICATION

*Rene Grzeszick, Leonard Rothacker, Gernot A. Fink*

TU Dortmund
Department of Computer Science
Email: {rene.grzeszick, leonard.rothacker, gernot.fink}@udo.edu

## ABSTRACT

This paper presents a novel method for combining local image features and spatial information for object classification tasks using the Bag-of-Features principle. The feature descriptor is extended by additional spatial information. Hence, similar feature descriptors do not only describe similar image patches, but similar patches in roughly the same region. Different spatial measures are evaluated on the Caltech101 dataset showing the improvement by incorporating spatial information into the feature descriptor. Furthermore, the method achieves better classification rates than the comparable Spatial Pyramids with lower a dimensional representation.

***Index Terms***— Image classification, Bag-of-Features, Spatial Pyramids

## 1. INTRODUCTION

Classifying visual objects is useful in many ways: The attention of visual systems can be drawn to interesting parts of an image, robots can learn about their environment, dangerous objects can be detected by surveillance systems, or information about the environment can be provided (e.g. in augmented reality applications). A major difficulty to the classification of objects is that they vary strongly in their appearance. There is no prior knowledge about their shape, color or very distinctive features that can be exploited.

In the last decade Bag-of-Features representations (cf. [1]) became very popular for image classification. Local image features describing the visual appearance of a small patch are extracted from a training set, clustered and quantized. Hence, a fixed set of representatives, the so-called *visual words*, are used for describing the features. An image is represented by the absolute or relative frequencies of the occuring visual words, the so-called *term-vector*. The presence or absence of different visual words in the term-vector indicates different object categories.

In object classification a major shortcoming of the Bag-of-Features representation is the lack of spatial information. Most objects in natural scenes have a similar orientation and a spatial structure that can be exploited, e.g. the wheels of a car are on the ground and below the chassis. Hence, the Bag-of-Features approach is often used in combination with *Spatial Pyramids* [2, 3, 4] which re-introduce a coarse representation of spatial information by subdividing the image and creating a Bag-of-Features for each sub-region. A concatenation of these term-vectors is used for describing the image. Spatial Pyramids have successfully been applied to object classification on various datasets like Caltech101 or VOC2007 [2, 5]. They were computed based on different local image descriptors, like SIFT [6, 7] or Centrist [8] and also used for tasks like scene categorization or object detection [8, 9]. Lately, these representations became increasingly high dimensional, using up to 8,000 visual words and 21 sub-regions which equals to 168,000 dimensions [5].

Such high dimensional representations are expensive in memory and computation time, especially for large sets of images. Manually labeled datasets like the VOC2011/2012 already contain about 30,000 images. To go even further, automatically obtained datasets can fastly grow into millions of sample images. In [10, 11] about 80 million images were collected and used for recognition tasks. Hence, a lower dimensional representation of the data is more than desirable.

We propose a method that incorporates spatial information at feature level. The appearance features that are extracted from the image are combined with spatial features. After clustering the spatial information is implicitly included in the visual vocabulary. For every spatial region only those representatives are stored that are observed in the training data, which allows for a lower dimensional representation of the data.

## 2. METHOD

The central idea of our method is to add spatial information to the local feature descriptors before they are clustered and quantized. We construct a new feature vector $\boldsymbol{v}$ consisting of an appearence feature vector $\boldsymbol{a}$ and a spatial feature vector $\boldsymbol{s}$:

$$\boldsymbol{v} = (a_0, ..., a_n, s_0, ..., s_n) \tag{1}$$

After clustering the visual words do not only represent local image descriptors that are similar in their appearance,



published in: Int. Conf. on Image Processing (ICIP) 2013

but similar appearance features in roughly the same spatial region. We refer to them as *Spatial Visual Words* and to the complete set as the *Spatial Visual Vocabulary*.

The proposed method is illustrated in Fig. 1. From a set of training images densly sampled SIFT features (cf. [2, 5, 12]) are extracted. Based on spatial quantization techniques, which are described in section 2.1 and 2.2, a spatial feature $\boldsymbol{s}$ is incoporated into the feature descriptor $\boldsymbol{v}$. Applying the Bag-of-Features principle, all modified descriptors are clustered in order to form a Spatial Visual Vocabulary. For clustering we use the generalized Lloyd algorithm (that is often referred to as K-Means; [13]). The Spatial Visual Vocabulary is used for quantizing the features of each training image and describing an image by a term-vector of Spatial Visual Words. These term-vectors are then used for training a Support Vector Machine (SVM). In the test case the appearance features are extracted from a single image, the spatial features are appended and the descriptors are then quantized with respect to the Spatial Visual Vocabulary from the training data. The SVM is used for predicting a class label for the test image.

### 2.1. XY representation

Spatial Pyramids subdivide an image in a quadtree-like manner, which implicitly uses the assumption that the visual object is roughly centered in the image. Using the same assumption, the direct translation of this approach to feature level would be adding quantized $xy$-coordinates as a spatial feature. In [2] it has been shown that the most important information of the Spatial Pyramid is contained within its top level. This representation can be approximated using Spatial Visual Words. For a subdivision with $2 \times 2$ subregions, $xy$-coordinates representing the upper-left, upper-right, lower-left or lower-right subregion are used. Hence, the new feature vector is described by:

$$\boldsymbol{v} = (a_0, a_1, ..., a_n, q(x), q(y)) \tag{2}$$

where $q(\cdot)$ denotes the quantization of the respective coordinate. For each of the regions represented by the spatial quantization the similar appearance features form clusters in the feature space, as shown in Fig. 1. Note that in order to achieve this behaviour the values of the $xy$-coordinates need to dominate the appearance features. For the SIFT features the 128 dimensional descriptor is divided by the average descriptor length, so that the sum of all dimensions becomes approximately one. Then $xy$-coordinates that are much larger than the SIFT descriptors' values are appended for the spatial feature $s$. For example, for a $2 \times 2$ subdivision the four subregions can be represented by $[(0,0),(0,1),(1,0),(1,1)]$. In our experiments we evaluated subdivisions for different grid sizes up to a continuous approach. In the last case the spatial regions in which similar appearance features are grouped are uncovered during the clustering process.

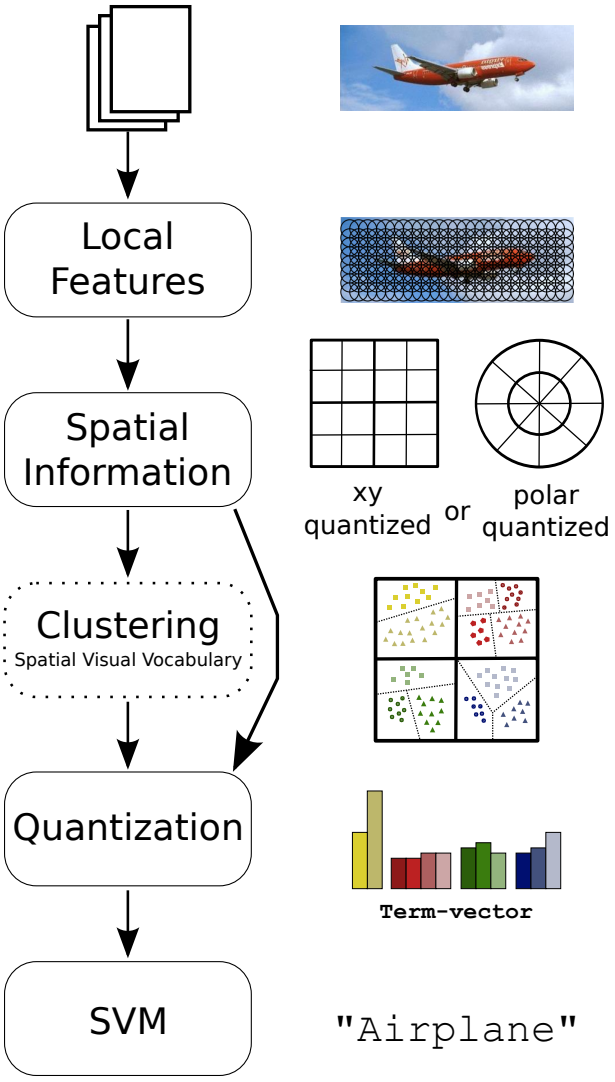


**Fig. 1**. Overview of the proposed method: Given an input image, local appearance features (e.g. SIFT) are extracted from the image based on a densly sampled grid. A spatial measure is used in order to combine the appearance features with spatial features. Common spatial features would be $xy$- or polar-coordinates. During the training the modified descriptors from all training images are clustered in order to form a Spatial Visual Vocabulary that holds the important information of each spatial region. The features of an image are quantized with respect to that vocabulary and represented by a set of Spatial Visual Words. An SVM is used for classification. This graphic is best viewed in color; the airplane image is taken from the Caltech101 database [14].

A subdivision into $n$ bins with a codebook size of $|V|$ using the Spatial Pyramid is approximated by quantizing $n$ $xy$-coordinates with $n * |V|$ spatial visual words. In both cases the final representation will have the dimension $n * |V|$. The only difference is that the pyramid uses the same codebook

for all subregions, while our codebooks are specific for each region. In our experiments we will show that very high dimensional codebooks contain redundant information and our approach allows to build smaller more specific codebooks for each region and therefore, is able to reduce the overall size of the representation.

### 2.2. Polar representation

Besides approximating the Spatial Pyramid approach, Spatial Visual Vocabularies can easily be adapted to different spatial quantization techniques. We also evaluated a polar coordinate representation that consists of two parts: 1. The distance toward the center of the image $r$. 2. The angle $\alpha$ to an upward axis. The spatial feature is then described by:

$$\boldsymbol{s} = (q(r), q(sin\ \alpha), q(cos\ \alpha)) \tag{3}$$

Again the SIFT descriptor is divided by the average descriptor length, so that it becomes approximately one. The radius $r$ is divided by the image diameter so that it becomes $0.5$ in the images corner points. In our experiments we evaluated a continuous case as well as quantizing all three values of the spatial feature, e.g. in $0.15$ or $0.25$ steps. This yields a representation in which the regions close to the center of the image are smaller than the regions in the corners of the image, which are more likely to show background clutter. Even smaller inner regions could be computed by using log-polar-coordinates. However, experiments showed that the regions become too small to contain enough information to create a statistical model, like the Bag-of-Features.

In our experiments we will show that the polar representation offers a significant advantage over subdiving the image in a quadtree like manner.

## 3. EVALUATION

We used the Caltech101 database [14] in order to evaluate our method. This database contains 101 different object categories. There are between 30 and 800 images per category, which sum up to 8677 images for all 101 categories. The images are mainly about 300x300 pixels in size and show only one visual object. Therefore, we do not have to consider the problem of detecting different visual objects in an image.

Following the experimental setup from Lazebnik, Schmid and Ponce that is described in [2], 30 images of each category are randomly chosen for training the classifier and densly sampled SIFT features with a step width of $8$ pixels and a descriptor size of $16 \times 16$ pixels are extracted. Note that all SIFT descriptors are computed with an upward orientation instead of rotating them in the direction of the main gradient. Hence, our descriptor modifications are directly comparable to the Spatial Pyramids.

| Method | Dim. | Classification rate |
|---|---|---|
| Bag-of-Features [2] | 200 | $41.2 \pm 1.3\%$ |
| 1-level Spatial Pyramid [2] | | |
| Top level | 800 | $55.9 \pm 0.9\%$ |
| Complete pyramid | 1,000 | $57.0 \pm 0.8\%$ |
| 2-level Spatial Pyramid [2] | | |
| Top level | 3,200 | $60.3 \pm 0.9\%$ |
| Complete pyramid | 4,200 | $64.6 \pm 0.7\%$ |
| Spatial Visual Vocabulary $2 \times 2$ regions | 200 | $52.5 \pm 0.9\%$ |
| | 500 | $56.1 \pm 0.9\%$ |
| | 800 | $57.4 \pm 0.9\%$ |
| | 1,000 | $58.1 \pm 0.9\%$ |
| | 2,000 | $60.6 \pm 0.9\%$ |
| | 4,000 | $59.9 \pm 0.9\%$ |
| Spatial Visual Vocabulary $4 \times 4$ regions | 1,000 | $63.4 \pm 0.8\%$ |
| | 2,000 | $63.9 \pm 0.8\%$ |
| | 3,200 | $64.7 \pm 0.8\%$ |
| | 4,200 | $64.0 \pm 0.8\%$ |

**Table 1**. Classification rates on the Caltech101 database. Spatial Visual Vocabularies are compared with a Bag-of-Features representation and Spatial Pyramids. The top level of the 1-level pyramid has $2 \times 2$ subregions and the 2-level pyramid $4 \times 4$ subregions. The pyramids are compared to Spatial Visual Vocabularies with $xy$-quantized coordinates that form similar subregions.

### 3.1. Comparison with Spatial Pyramids

In the first experiments, which are shown in Table 1, we evaluated models that are comparable to the top level of Spatial Pyramids. The results for complete pyramids using a vocabulary size of $|V| = 200$ were reproduced with our pipeline with 56.6% for one level and 64.0% for two levels. The total dimension of these equals to $1,000$ and $4,200$ respectively. In comparison with the 1-level Spatial Pyramid we used Spatial Visual Words with quantized $xy$-coordinates and $2 \times 2$ subregions. Note that the Spatial Visual Vocabulary achieves significantly better results than the top level of the Spatial Pyramid. Also, using the overall same dimensionality it is significantly better than the complete pyramid. This shows that specific codebooks, computed for each subregion, represent the problem domain better than one common codebook. Concerning the parameterization of the quantized $xy$-values we experimented with $(0.25, 0.75)$, $(0, 1)$ and $(0, 5)$ coordinates for the subregions. Our experiments showed that there is no significant difference, as long as the spatial feature dominates the appearance feature.

When computing smaller codebooks our experiments show that Spatial Visual Vocabularies yield the same results as the top level of the pyramid using only 500 dimensions instead of 800. It is also crucial that we are able to choose the dimension in a much finer manner than the original pyramid.

Removing one visual word from a Spatial Visual Vocabulary reduces the dimension by one, whereas the removal of one Visual Word in the Spatial Pyramid affects all subregions.

The importance of spatial information for object classification is shown by the comparison with a Bag-of-Features representation. Using $|V| = 200$ visual words the representation using Spatial Visual Vocabularies achieves a classification rate of 52.5%, which outperforms the Bag-of-Features approach.

Comparable results can be observed for the 2-level pyramid. Using Spatial Visual Words that are quantized into $4 \times 4$ subregions our model achieves significantly better results than the top level of the respective Spatial Pyramid. When using a vocabulary size of $|V| = 1,000$ the classification rates are better, while the dimensionality is smaller than a third. Further increasing the dimensionality to $3,200$ and $4,200$ visual words on the other hand does not show a large improvement. At $|V| = 3,200$ the results are as good as the complete pyramid, but the comparably small difference shows that the subdivision into $4 \times 4$ subregions does not allow to further improve the classification rates by simply increasing the size of the visual vocabulary. This is also confirmed by the even larger vocabulary size of $|V| = 4,200$, which does not increase the classification rate anymore.

While our method is a significant improvement compared to the 1-level pyramid from [2] and yields results that are comparable to a 2-level pyramid with much lower dimensionality, the state-of-the-art results using Spatial Pyramids are still significantly better. In [5] classification rates of up to $76.9\%$ are reported. The are mainly three reasons for this: 1. More than $70\times$ more SIFT features are extracted from the images. 2. Complex feature encoding techniques, like Locality-constrained linear coding [15], are applied. 3. Higher dimensional vocabularies and more pyramid bins are used. This leads to a increased dimensionality of 168,000, which we tried to avoid in our work. However, all of these modifications can also be applied to Spatial Visual Vocabularies. Especially the sampling of more SIFT features and using feature encoding techniques might be of further interest, since they do not increase the dimensionality of the representation.

### 3.2. Spatial configurations

In further experiments we compared different levels of detail for the quantized spatial coordinates from coarse subregions to a continuous approach. In the continuous case we appended the exact position of a visual word. We used different normalization ranges, having the best results with $[0, 1]$ for $xy$-coordinates and $[0, 0.5]$ for the radius. However, as shown in Table 2, the quantization of the spatial coordinates adds some level of abstraction to the spatial information that is very important for the classification. When appending continuous values the clustering is not able to uncover a set of regions that is able to perform as good as the quantized ones.

| Method | Classification rate |
|---|---|
| XY quantized - $2 \times 2$ cells | $58.1 \pm 0.9\%$ |
| XY quantized - $4 \times 4$ cells | $63.4 \pm 0.8\%$ |
| XY - continuous | $55.0 \pm 0.9\%$ |
| Polar quantized - .50 steps | $63.7 \pm 0.8\%$ |
| Polar quantized - .25 steps | $64.9 \pm 0.8\%$ |
| Polar quantized - .15 steps | $64.5 \pm 0.8\%$ |
| Polar - continuous | $59.4 \pm 0.9\%$ |

**Table 2**. Classification rates on the Caltech101 database using Spatial Visual Vocabularies with $|V| = 1,000$ dimensions. Different spatial quantization techniques from coarse to continuous coordinates are evaluated. Also, the $xy$-quantized representation is compared with a polar representation.

When comparing the $xy$-coordinates with the polar coordinates, our results show that the polar representation yields a significant improvement. There are two aspects to this: First, the polar subregions are finer than the $2 \times 2$ quantized $xy$ values. Nevertheless, the polar measures also outperform the $xy$-configuration using $4 \times 4$ regions that have a comparable level of detail to .25 quantization steps in the polar representation. More importantly the polar measure forms finer regions at the center of the image and coarser at the borders of it. It is much more likely that the border regions carry background information and are therefore coarser in the spatial subdivision of the image.

Note however that this approach uses the assumption that the dominant object is roughly centered, which is true for most real life images, but may not hold for very complex scenes. Nevertheless, our approach would also allow to move the reference point from the center of an image toward a different region of interest. For example, a proto-object detector as shown in [16] could be used.

## 4. CONCLUSION

We presented a novel method for incorporating spatial information into the Bag-of-Features representation at feature level. Our experiments on the Caltech101 dataset show that the lack of spatial information is a crucial disadvantage and that encoding this information at feature level yields a significant improvement.

Our method is comparable to the top level of Spatial Pyramids, but computes a Spatial Visual Vocabulary that is specific for each subregion. By generating such Spatial Visual Vocabularies our method outperforms the results from the literature that use the same features and comparable dimensionality. Furthermore, by using quantized polar coordinates, instead of a quadtree-like subdivision of the image, the results could be improved even further.

## 5. REFERENCES


[1] S. O'Hara and B.A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *Arxiv preprint arXiv:1101.3354*, 2011.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, vol. 2, pp. 2169–2178.

[3] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 401–408.

[4] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *IEEE International Conference on Computer Vision (ICCV)*, 2005, vol. 2, pp. 1458–1465.

[5] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference*, 2011.

[6] D.G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision (ICCV)*, 1999, vol. 2, pp. 1150–1157.

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] J. Wu and J.M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[9] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 1585–1592.

[10] A. Torralba, R. Fergus, and W.T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.

[11] R. Fergus, Y. Weiss, and A. Torralba, "Semi-supervised learning in gigantic image collections," *Neural Information Processing Systems*, 2009.

[12] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, vol. 2, pp. 524–531.

[13] Stuart Lloyd, "Least squares quantization in PCM," *Information Theory, IEEE Transactions on*, vol. 28, no. 2, pp. 129–137, 1982.

[14] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories.," in *CVPR 2004, Workshop on Generative-Model Based Vision*. IEEE, 2004.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR 2010*. IEEE, 2010, pp. 3360–3367.

[16] Fabian Nasse and Gernot A. Fink, "A bottom-up approach for learning visual object detection models from unreliable sources," in *Pattern Recognition: 34th DAGM-Symposium Graz*. 2012, pp. 428–435, Springer.