

Spatial Extensions to Bag of Visual Words

Ville Viitaniemi
Dept. of Information and Computer Science
Helsinki University of Technology (TKK)
P.O. Box 5400, 02015 TKK, Finland
ville.viitaniemi@tkk.fi

Jorma Laaksonen
Dept. of Information and Computer Science
Helsinki University of Technology (TKK)
P.O. Box 5400, 02015 TKK, Finland
jorma.laaksonen@tkk.fi

ABSTRACT

The Bag of Visual Words (BoV) paradigm has successfully been applied to image content analysis tasks such as image classification and object detection. The basic BoV approach overlooks spatial descriptor distribution within images. Here we describe spatial extensions to BoV and experimentally compare them in the VOC2007 benchmark image category detection task. In particular, we compare two ways for tiling images geometrically: soft tiling approach—proposed here—and the traditional hard tiling technique. The experiments also address two methods of fusing information from several tilings of the images: post-classifier fusion and fusion on the level of a SVM kernel.

The experiments confirm that the performance of a BoV system can be greatly enhanced by taking the descriptors' spatial distribution into account. The soft tiling technique performs well even with a single tiling mask, whereas multi-mask fusion is necessary for good category detection performance in case of hard tiling. The evaluated fusion mechanisms performed approximately equally well.

Categories and Subject Descriptors

I.4.8 [Image Processing and Computer Vision]: Scene Analysis—*Object recognition*

General Terms

Measurement Performance

Keywords

local descriptors, image category detection, bag of visual words

1. INTRODUCTION

In supervised image category detection the goal is to predict whether a novel test image belongs to a category defined by a training set of positive and negative example images. The categories can correspond, for example, to the presence

or absence of a certain object, such as a dog. In order to automatically perform such a task based on the visual properties of the images, one must use a representation for the properties that can be extracted automatically from the images.

Histograms of local features have proven to be powerful image representations for image classification and object detection. Consequently their use has lately become commonplace in image content analysis tasks (e.g. [11, 18]). This paradigm is also known by the name Bag of Visual Words (BoV) in analogy with the successful Bag of Words paradigm in text retrieval. In this analogue, images correspond to documents and different local descriptor values to words.

Use of local image feature histograms for supervised image classification and characterisation can be divided into several steps:

1. Selecting image locations of interest.
2. Describing each location with suitable visual descriptors (e.g. SIFT).
3. Characterising the distribution of the descriptors within each image with a histogram.
4. Using the histograms as feature vectors representing the images in a supervised vector space algorithm, such as SVM.

Figure 1 schematically shows these steps of the BoV pipeline.

In its basic form the BoV approach loses all information about the interest points' spatial distribution within images. However, many image categories are such that such spatial structure could be useful in their detection. A common extension to BoV is to geometrically partition all the images with the same tiling pattern. Each part is then described with a separate histogram and the image dissimilarity is formulated as the sum of the dissimilarities of corresponding tiles. A further extension that makes the approach significantly more useful is to combine the information obtained by the use of several different tiling masks, such as in the spatial pyramid technique of [6].

In this paper we experimentally compare various spatial extensions to BoV in a concrete image category detection task defined in the VOC2007 benchmark. We partition the images with various alternative geometric masks. Here we denote this traditional technique as *hard tiling*. We also introduce the technique of *soft tiling*, where each interest point is assigned simultaneously to several spatial tiles, to each in varying degree. This is analogous to the soft histogram technique in codebook space that we have found to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIVR '09 July 8-10, 2009 Santorini, GR

Copyright 2009 ACM 978-1-60558-480-5/09/07 ...\$5.00.

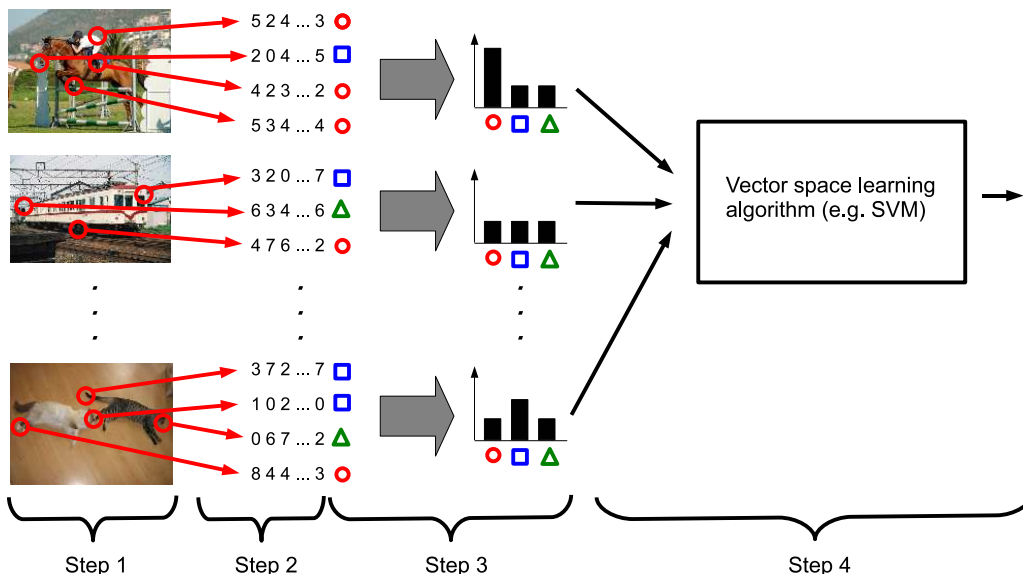


Figure 1: Steps in the supervised BoV pipeline

be useful in [15]. Furthermore, we consider fusing information from various combinations of masks—corresponding to different resolutions—and histograms of different sizes with which the individual tiles are described. For the fusion we employ two alternative techniques: multi-scale kernels (simplified version of [8]) and post-classifier fusion by Bayesian logistic regression.

The rest of this paper is organised as follows. In Sect. 2 we outline our BoV implementation. The considered spatial extensions to BoV are described in Sect. 3. In Sect. 4 we first detail the image category detection task and experimental procedures we subsequently use for experimentally comparing the spatial techniques in the rest of the section. In Sect. 5 we summarise the paper by final conclusions.

2. BASELINE BOV SYSTEM

In the first stage of our implementation of the BoV pipeline, interest points are detected from each image with a combined Harris-Laplace detector [9] that outputs around 1200 interest points per image on average with the images used in the current experiments. In step 2 the image area around each interest point is individually described with a 128-dimensional SIFT descriptor [7], a widely-used and rather well-performing descriptor. In step 3 each image is described by forming a histogram of the SIFT descriptors. We determine the histogram bins by clustering a sample of the interest point SIFT descriptors (20 per image) with the Linde-Buzo-Gray (LBG) algorithm. In our earlier experiments [14] we have found such codebooks to perform reasonably well while the computational cost associated with the clustering still remains manageable. In our system we use histograms with sizes ranging from 128 to 16384.

In the final fourth step the descriptor histograms of both training and test images are fed into supervised classifiers, separately for each of the 20 object classes. We use weighted C-SVC variants of the SVM algorithm, implemented in the version 2.84 of the software package LIBSVM [3]. As the

kernel function g we use the exponential χ^2 -kernel

$$g_{\chi^2}(\mathbf{x}, \mathbf{x}') = \exp(-\gamma d_{\chi^2}(\mathbf{x}, \mathbf{x}')) \quad (1)$$

with the χ^2 distance given by

$$d_{\chi^2}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d \frac{(x_i - x'_i)^2}{x_i + x'_i}. \quad (2)$$

The free parameters of the classifiers are chosen by a search procedure maximising cross-validation performance in the training set. Details of the SVM classification stage can be found in [13]. The first row of Table 1 summarises the performance of the basic BoV implementation for different codebook sizes in the image category detection task of Sect. 4.1.

In some techniques we propose to fuse together several different histograms. To provide comparison reference for these techniques, we consider the performance of post-classifier fusion of the detection results of the histograms in question. For classifier fusion we employ Bayesian Logistic Regression (BBR) [1] that that we have found usually to perform at least as well as the other methods we have evaluated (SVM, sum and product fusion mechanism) for small sets of similar features.

2.1 Speed-up Technique

For the largest codebooks, describing images with histograms becomes impractically time-consuming if implemented in a straightforward fashion. This is because determining the nearest histogram bin is an operation linear in the codebook size. Therefore, a speed-up structure is employed to facilitate fast approximate nearest neighbour search. Such structures have been used also earlier, e.g. in [10]. There the authors ended up using the hierarchical speed-up mechanism both in codebook training and nearest neighbour search. In our approach we train the codebooks without approximations. This is naturally more time-consuming, but provides somewhat more accurate results.

Our speed up-structure for a given trained codebook is formed by creating a succession of codebooks diminishing in size with the k-means clustering algorithm. First invocation of the algorithm takes the N original codebook vectors as input and partitions them into clusters whose number is N divided by a shrinking factor s . The resulting cluster centers are used as input to the next invocation of k-means. This is repeated until the clustering produces only a few cluster centers.

The structure is employed in the nearest-neighbour search of vector \mathbf{v} by first determining the closest match of \mathbf{v} in the smallest of the codebooks. Then \mathbf{v} 's approximate best match is located in the next larger codebook. This is done by limiting the search to n_u codebook vectors of the larger codebook that are closest to the codebook vector of the smaller codebook. This way a match is found in successively larger codebooks, and eventually among the original codebook vectors. The last search is performed among a different number n_b of codebook vectors. The time cost of this search algorithm is proportional to the logarithm of the codebook size (for constant n_u and n_b). According to our experiments, the approximative algorithm is nearly as accurate as the full search in terms of mean square quantisation error.

2.2 Soft Histogram Technique

Histograms with different number of bins encode the visual information in descriptors of interest points with different granularities. In another study [15] we have compared some alternative methods for synthesizing histogram-like feature vectors that combine the information from different levels of granularity. The most effective approach among the investigated ones was found to be the *soft histogram* technique that was proposed also in [12]. In *hard* clustering, the descriptor of each interest point is assigned to exactly one histogram bin, the bin found by the nearest neighbour search. In contrast, in the soft histograms the increment is distributed among several neighbouring histogram bins, in proportion to their similarity to the interest point descriptor.

Table 1 demonstrates the performance gain due to the soft histogram technique that was combined with the speedup technique. It is worth mentioning that one of the other methods investigated in [15] was the descriptor-space analogue of the spatial multi-scale kernel method of Sect. 3.2, closely related to the pyramid match kernel of [5]. The performance gain over single granularity due to this multi-granularity kernel method was less than half the gain brought by the soft histogram method. As we considered the gain due to soft histograms to be substantial, the technique was applied in all the experiments comparing different spatial techniques. The compability of the soft histogram technique with the evaluated spatial techniques was later confirmed with a separate set of experiments (Section 4.4).

3. SPATIAL TECHNIQUES

In this section we describe methods for exploiting the spatial information in the interest point descriptors. We consider doing this by tiling the image area into several tiles, forming separate histograms for each tile and then concatenating the resulting histograms. The tilings—schematically shown in Fig. 2—divide the image into 2×2 , 3×3 , 4×4 , 5×5 or 6×6 rectangular tilings. We consider also 5-part (cs-5) and 10-part (cs-10) center-surround tilings. The remainder

of this section details two techniques that can be used as components in the previously outlined framework for using spatial information.

3.1 Soft Tiling

Analogously to the soft histogram bin determination in connection with image-wide global histograms (Sect. 2.2), we can assign the interest points not only to one spatial image tile (traditional *hard tiling*) but several ones with varying degrees (*soft tiling*). The soft tiling can be presented with spatially varying tile membership masks. Here we have normalised the memberships of each image pixel to sum to one.

Figure 3 shows some membership masks that are smoothed versions of the rectangular 2×2 , 4×4 and the center-surround tilings. The dark areas of the images correspond to large membership degrees. The left side of the figure shows the membership masks of the 2×2 tiling and the masks of the diagonal tiles of the 4×4 tiling. On the right some masks of the 5-part and 10-part center-surround tilings are shown. The remaining center-surround masks follow by symmetry operations.

The tiling masks have been devised to resemble the corresponding hard tiling masks. The rectangular tilings were smoothed by placing a Gaussian in the center of each tile. The 5-part tiling was as a radial function modulated by an angular part (except for the center tile). We used logistic sigmoid as the radial function and a Gaussian as the angular one.

3.2 Multi-scale Kernel

In this technique the kernel of the SVM is extended to take into account not only a single SIFT histogram H , but a whole set of histograms $\{H_i\}$, H_i being the concatenation of histograms describing the tiles of resolution i . To form the kernel, we evaluate the multi-scale distance d_{ms} between two images as a weighted sum of distances d_i in different spatial resolutions i :

$$d_{ms} = \sum_i w_i d_i, \quad w_i = N_i^{1/K}, \quad (3)$$

where d_i is the χ^2 -distance between histogram concatenations resulting from tiling i of N_i tiles and w_i is the corresponding weighting factor. Here K is a free parameter of the method that can be thought to correspond to the dimensionality of the space the histograms quantise. Value $K = \infty$ corresponds to unweighted concatenation of the histograms. In most of the experiments we evaluate two weighting variants, namely $K = \infty$ (flat weighting) and $K = 2$ (corresponding to the dimensionality of the image grid). We do not make a serious attempt to devise a method for selecting the value of K from the data. The distances are used to form a kernel for the SVM by

$$g_{ms} = \exp(-\gamma d_{ms}). \quad (4)$$

This multi-scale kernel method is essentially the same as the one of [8] with the exception of the simpler form of weighting the resolutions employed here. In [8], the weights were learned from the training data. Both these methods largely resemble the spatial pyramid method of [6] that adapts pyramid matching method of [5] to spatial domain. Also there the image similarity is a weighted sum of similarities of histograms of different spatial resolutions. However,

Table 1: MAP performance of the baseline BoV system for histograms of different sizes in the VOC2007 image category detection task detailed in Sect. 4.1, both with and without the use of the soft histogram technique

	128	256	512	1024	2048	4096	8192	16384
baseline	0.357	0.376	0.387	0.397	0.400	0.404	0.398	-
soft	-	0.385	0.406	0.419	0.435	0.438	0.448	0.451

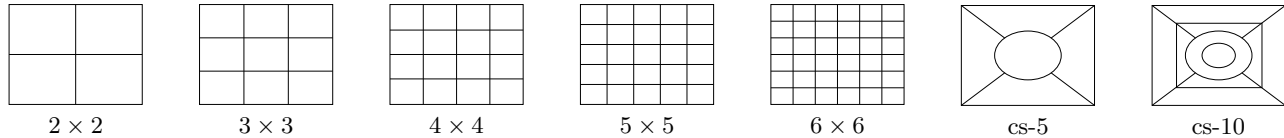


Figure 2: Tiling patterns for partitioning the image area.

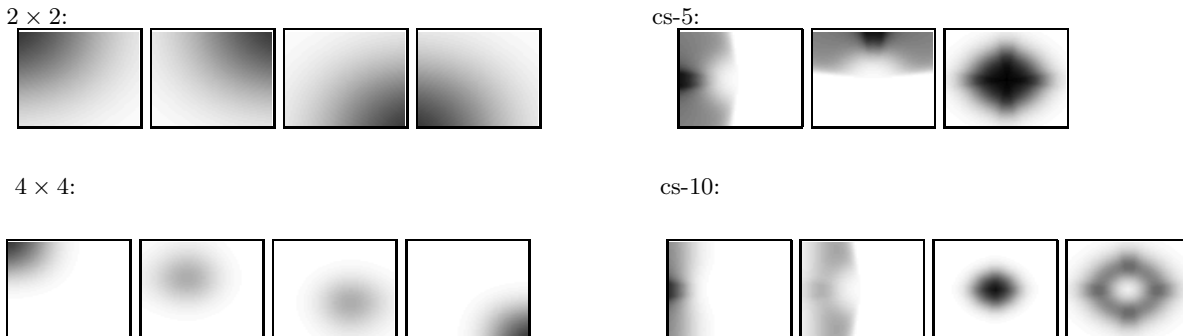


Figure 3: Membership masks of some spatially smooth tilings

they use histogram intersection as the similarity measure. Experiments (e.g. [2]) show that an exponential χ^2 -kernel reflects histogram similarity better than a kernel based on histogram intersection in the present application.

4. EXPERIMENTS

4.1 Category Detection Task and Experimental Procedures

In the experiments we consider the supervised image category detection problem. Specifically, we measure the performance of several algorithmic variants for the task using images and categories defined in the PASCAL NoE Visual Object Classes (VOC) Challenge 2007 collection [4]. In the collection there are altogether 9963 photographic images of natural scenes. In the experiments we use the half of them (5011 images) denoted “trainval” by the challenge organisers. Each of the images contains at least one occurrence of the predefined 20 object classes, detailed in Table 2, including e.g. several types of vehicles, animals and furniture. The presences of these objects in the images were manually annotated by the organisers. Figure 4 shows some examples of the images and objects. In many images there are objects of several classes present. In the experiments (and in the “classification task” of VOC Challenge) each object class is taken to define an image category.

In the experiments the 5011 images are partitioned approximately equally into training and test sets. Every experiment is performed separately for each of the 20 object classes. The category detection accuracy is measured in

Table 2: The 20 object classes of VOC Challenge 2007

aeroplane	bus	dining table	potted plant
bicycle	car	dog	sheep
bird	cat	horse	sofa
boat	chair	motorbike	train
bottle	cow	person	tv/monitor

terms of non-interpolated average precision (AP). The AP values were averaged over the 20 object classes and six different train/test partitionings. The average MAP values tabulated in the result tables had 95% confidence intervals of order 0.01 in all the experiments. This means that for some pairs of techniques with nearly the same MAP values, the order of superiority can not be stated very confidently on basis of a single experiment. However, in the experiments the discussed techniques are usually evaluated with several different histogram codebook sizes and other algorithmic variations. Such experiment series usually lead to rather definitive conclusions. Moreover, because of systematic differences between the six trials, the confidence intervals arguably underestimate the reliability of the results for the purpose of comparing various techniques. The variability being similar for all the results, we do not clutter the tables of results with confidence intervals

4.2 Individual Tilings

In the experiments of this section we compare the im-



Figure 4: Examples of VOC Challenge 2007 images and their annotations

age category detection performances the hard and soft tiling techniques achieve when the image area is partitioned with a single geometrical tiling mask. In Table 3 we show the MAP performances for both techniques, and several different tiling masks. For the table we described each tile with 2048 bin soft histograms. Figure 5 illustrates the same results.

It can be noticed that soft tiling results in significantly larger MAP values. Moreover, whereas the MAP for the hard tilings start to diminish for finer tilings, the same effect is only barely noticeable for the soft tilings. This could be attributed to the soft tiling technique effectively integrating information from coarser resolutions into finer ones. Table 3 also shows that each of the individual soft tilings outperforms the corresponding image-wide (global) histogram by a wide margin. For the hard tilings, this is not the case.

We also performed similar experiments with a range of histogram sizes. Similar observations as made above in the case of 2048-bin histograms can be made also from the results of these experiments, detailed in Table 4. Each entry of the table shows performance of hard tiling (h) followed by that of soft tiling (s). Generally, we obtained our best results with the largest codebooks we tried, i.e. 8192 for the coarser tilings. Earlier it has been reported [16] that finer tilings would be best described with smaller codebooks (i.e. histogram granularities). We surmise the explanation to lie—in analogy with the spatially soft tiling—in the granularity-integrating nature of the soft histogram technique. With these larger codebooks, the highest MAP obtained with the soft tiling technique was 0.476, shared by cs-5 and cs-10 tilings (other tilings not far behind). For the hard tiling, the highest MAP was 0.443.

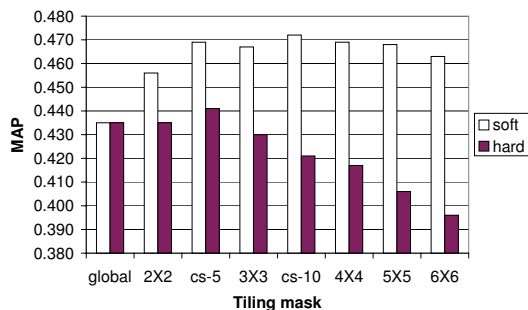


Figure 5: MAP performances of the spatially hard and soft tiling techniques with codebook size 2048

4.3 Fusion of Tilings

In the previous section we clearly saw the spatially soft tiling to be more accurate when image area is partitioned with a single tiling mask. However, in practice there is no need to limit oneself to use a single tiling mask, but information from several tilings can be fused together. In the experiments of this section, we consider two fusion approaches: the multi-scale kernel technique of Sect. 3.2 and post-classifier fusion based on BBR.

For the across-scale fusion one needs decide 1) which set of tilings one combines, and 2) with how many histogram bins each tile is described. There is no clear practical guideline to answer these questions. We thus experimented with numerous alternatives. In Table 5 we report the category detection MAP for some of them when information from different tilings have been combined with both the multi-scale kernel technique and post-classifier BBR fusion.

In the table each pair of rows corresponds to a certain combination of tilings. On rows labeled “hard” and “soft”, the histograms have been collected employing either hard or soft tiling, respectively. For each tiling combination, the included tilings are shown in the “Contents” field of the table. In this field there is a column for each tiling. If the tiling is included in the combination the number in the column indicates the size of histograms used to describe the tiles of that tiling. The four rightmost columns of the table list the MAP performances achieved by combining the tilings. The columns “ $K = 0$ ” and “ $K = \infty$ ” display the MAP values obtained by the multi-scale kernel technique. The columns “BBR” and “indiv.” are included for comparison. The column “BBR” shows the MAP value obtained by the post-classifier BBR fusion of the resolutions. Column “indiv.” indicates the category detection MAP achieved using the best one of the individual tilings alone.

What comes to selection of tilings to be fused, our conclusion from the experiments is rather vague: it seems useful to let the tilings have some redundancy, but not too much. Of the individual tilings, the center-surround tilings seem to perform rather well in comparison with the rectangular tilings.

On the issue of describing the tiles of a single resolution tiling we can be more definite: in our experiments with soft histograms it was useful to describe each tiling with the largest of the considered codebooks. In practice, this means that the tiles of coarser partitionings—such as 2×2 — can be described with larger histograms than the tiles of finer resolutions (e.g. 6×6), as the storage requirements for full-size histograms with the finer tiling masks may become pro-

Table 3: MAP performances of the spatially hard and soft tiling techniques with codebook size 2048

spatial tiling	tiling								
	global	2 × 2	cs-5	3 × 3	cs-10	4 × 4	5 × 5	6 × 6	
hard	0.435	0.435	0.441	0.430	0.421	0.417	0.406	0.396	
soft	0.435	0.456	0.469	0.467	0.472	0.469	0.468	0.463	

Table 4: MAP performance with a larger variety of soft histograms employing both hard (h) and soft (t) spatial tiling

	global	tiling		cs-5	3 × 3	cs-10	4 × 4	5 × 5	6 × 6
		2 × 2	h/s						
bins	512	0.406	0.419/0.433	—/0.452	0.422/0.446	0.409/0.453	—/0.453	0.400/0.453	0.391/0.451
/tile	1024	0.419	0.430/0.449	0.436/0.462	0.430/0.458	0.419/0.463	0.415/0.462	0.405/0.462	0.397/0.458
	2048	0.435	0.435/0.456	0.441/0.469	0.432/0.467	0.421/0.472	0.417/0.469	0.406/0.468	0.396/0.463
	4096	0.438	0.437/0.457	0.439/0.472	0.431/0.470	0.421/0.472	0.416/0.472	—/—	—/—
	8192	0.448	0.440/0.466	0.443/0.476	0.432/0.475	0.396/0.476	—/—	—/—	—/—

Table 5: Combining information from several resolutions with either multi-scale kernel or BBR fusion mechanisms

Contents	global	tiling						MS kernel		BBR	indiv.	
		2 × 2	cs-5	3 × 3	cs-10	4 × 4	5 × 5	6 × 6	$K = \infty$			$K = 2$
128	128					128						
							hard	0.411	0.405	0.407	0.384	
							soft	0.408	0.391	0.415	0.417	
	128	128	128		128	128	hard	0.415	0.416	0.412	0.390	
							soft	0.419	0.404	0.420	0.420	
2048	128				128		hard	0.437	0.449	0.450	0.429	
							soft	0.443	0.447	0.452	0.429	
2048	128		128		128	128	hard	0.431	0.450	0.451	0.429	
							soft	0.439	0.448	0.457	0.429	
2048	512		512		128	128	hard	0.445	0.462	0.457	0.429	
							soft	0.457	0.461	0.459	0.447	
16384		512				512	hard	0.459	0.469	0.465	0.448	
							soft	0.470	0.472	0.474	0.451	
16384	4096	8192	2048	2048	2048	1024	512	soft	0.477	0.474	0.480	0.476
16384		8192		2048				soft	0.479	0.471	0.480	0.476
16384		8192		2048		1024		soft	0.481	0.472	0.483	0.476
		8192		2048				soft	0.476	0.476	0.478	0.476

hibitive. For comparison, in the spatial pyramid approach of [6] all the tiles of different resolutions are described with a histogram of the same size.

By comparing Tables 3, 4 and 5, it can be seen that when using the spatially soft tiling, the performances of the individual tilings come much closer to performances of the combination of tilings than by using hard tiling. There probably is much more across-scale redundancy in the histograms of the soft tiling technique. What hints to this direction is that the BBR fusion performance of the hard and soft tiling is not very different, although the soft tiling seems to consistently perform somewhat better. This point remains somewhat vague as we do not directly measure the redundancy of the classification scores of different tilings.

In general the multi-scale kernel often slightly outperforms the post-classifier fusion by BBR when the tilings are described with small histograms. The situation appears to be the opposite when tilings are described with larger histograms and the overall MAP is consequently higher. Yet, the differences are not large. Both resolution-fusing techniques clearly outperform the best of the individual resolu-

tions for small histograms. However, for larger histograms, the performance is not that greatly improved by fusing several resolutions. For comparison, [17] reported a slight advantage of SVM-based fusion over multi-scale technique of [6] for rather small tiling combinations. We tried also SVM-based fusion for some of the combinations, but BBR produced better results here.

The soft tiling technique does not seem to be as compatible with the multi-scale kernel technique as the hard tiling. Especially the weighting of resolutions seems to be problematic. For hard tiling, the weighting parameter $K = 2$ often brings notable improvements over the unweighted case ($K = \infty$), whereas for soft tiling the advantage is often only slight. This impression is backed up by Figure 6 which shows the results of a more extensive evaluation of the weight parameter K for one combination of tilings. For hard tiling, we observe a clear peak of performance around $K = 2$, whereas for soft tiling a broad range of weight parameters K resulted in similar performance.

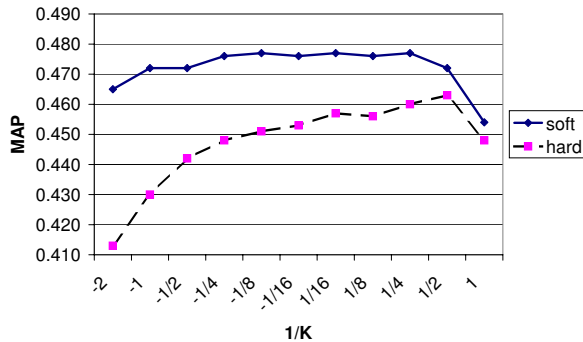


Figure 6: The effect of weight parameter K on the MAP performance of the multi-scale kernel technique for both hard and soft spatial tiling

4.4 Combining Spatial Techniques with Soft Histograms

In the above experiments we have combined information from several spatial scales, each scale being described with histograms that have been smoothed in SIFT space with the soft histogram technique. As mentioned in Section 2.2, this technique has been found to be useful as such. It can still be asked whether it is useful to combine the SIFT space soft histogram technique with spatial multi-resolution techniques, as we have thus far assumed. The results in Table 6 indicate this to be indeed the case. The soft histogram technique significantly improves the performance of spatial multi-scale combinations, both with soft and hard spatial tiling.

5. CONCLUSIONS

In this paper we have described and evaluated spatial local feature histogram techniques for the purpose of image category detection. The experiments confirm that the performance of a BoV system can be greatly enhanced by taking the descriptors' spatial distribution into account. The component techniques we have addressed are the partitioning the images with geometric tiling masks, and the fusion of the information obtained employing several different tiling masks.

The techniques can be seen as analogues of the techniques we have investigated earlier [15] for combining information from different descriptor space granularities. The spatial techniques discussed here are otherwise similar, but deal with information on several spatial scales. According to our observations, the two different multi-scale spaces are orthogonal in the way that the techniques can be usefully combined. However, even though the two scale spaces are analogous in the sense that they can be addressed with techniques based on similar ideas, they may still play a different role on conjunction of visual tasks. The spatial relationships in the 2D image plane have natural semantic meanings, whereas the descriptor feature space is an artificial construction without such straightforward semantic interpretations.

Of the tiling techniques, spatially soft tiling appears to be a useful technique even in the case of a single tiling mask. For applications where ultimate accuracy is not necessary, the image category could possibly be determined on basis of

a single soft tiling mask. This can be contrasted with the use of spatially hard tilings where fusion of several resolutions is an essential prerequisite for good performance.

By fusing information from several spatial scales, the performance of a system employing hard tiling greatly improves, whereas in the case of soft tiling the improvement is only minor. However, even when performing multi-scale fusion, the soft tiling technique provides some performance advantage over hard tiling. In the experiments, we did not observe essential differences in performances of multi-scale kernel and post-classifier BBR fusion methods. This might be an inevitable result as early and late fusion mechanisms have not generally been found to be superior to one another.

The performance gain obtained by fusing soft tilings of multiple scales makes the situation different from the one we encountered in conjunction with the descriptor-space techniques in [15]. There the further fusion of analogous soft histograms only degraded performance. This might be a manifestation of differences of the two scale spaces. Alternatively, this might be an indication of the form of the tile membership masks being suboptimal. This would be no surprise as the mask shapes were not optimised in any way. Specific issues that should be addressed are the overlap of the tiles, sharpness of the borders and form of the decay.

6. ACKNOWLEDGMENTS

Supported by the Academy of Finland in the *Finnish Centre of Excellence in Adaptive Informatics Research* project. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under *grant agreement n° 216529*, Personal Information Navigator Adapting Through Viewing, PinView.

7. REFERENCES

- [1] D. M. A. Genkin, D. D. Lewis. BBR: Bayesian logistic regression software, 2005. Software available at <http://www.stat.rutgers.edu/~madigan/BBR/>.
- [2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of ACM ICVR 2007*, pages 401–408, 2007.
- [3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [5] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 2007.
- [6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE CVPR*, volume 2, pages 2169–2178, 2006.
- [7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.
- [8] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. Presentation slides in

Table 6: Comparing the performance based on histograms employing either hard or soft assignment in SIFT space. The layout of the table is similar to that of Table 5 except that the leftmost column indicates the type of SIFT space bin assignment used when collecting the histograms

SIFT assignment	Contents				tiling	MS kernel		BBR	indiv.
	global	cs-5	cs-10	5×5		$K = \infty$	$K = 2$		
hard	4096	4096	2048	1024	hard	0.428	0.429	0.426	0.404
					soft	0.447	0.437	0.444	0.434
soft	4096	4096	2048	1024	hard	0.455	0.461	0.465	0.439
					soft	0.475	0.474	0.478	0.472

VOC2007 Workshop, 2007. Available via <http://lear.inrialpes.fr/pubs/2007/MSHV07/MarszalekSchmid-VOC07-LearningRepresentations-slides.pdf>.

- [9] K. Mikolajczyk and C. Schmid. Scale and affine point invariant interest point detectors. *International Journal of Computer Vision*, 60(1):68–86, 2004.
- [10] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE CVPR 2006*, volume 2, pages 2161–2168, 2006.
- [11] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. of ICCV'03*, volume 2, pages 1470–1477, Oct. 2003.
- [12] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of ECCV 2008*, pages 696–709, 2008.
- [13] V. Viitaniemi and J. Laaksonen. Improving the accuracy of global feature fusion based image categorisation. In *Proc. of SAMT 2007*, volume 4669 of *LNCS*, pages 1–14, Genova, Italy, December 2007. Springer.
- [14] V. Viitaniemi and J. Laaksonen. Experiments on selection of codebooks for local image feature histograms. In *Proc. of VISUAL2008*, volume 5188 of *LNCS*, pages 126–137. Springer, 2008.
- [15] V. Viitaniemi and J. Laaksonen. Combining local feature histograms of different granularities. 2009. Accepted to SCIA 2009.
- [16] J. Yang, Y.-G. Jiang, A. G.Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proc. of MIR '07*, pages 197–206, 2007.
- [17] J. Zhang. Local features and kernels for classification of object categories. Presentation slides in VOC2006 Workshop, 2006. Available via <http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006/workshop.html>.
- [18] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.