Multi-level local descriptor quantization for Bag-of-Visterms image representation

Pedro Quelhas
IDIAP Research Institute
Martigny, Switzerland
pedro.quelhas@gmail.com

Jean-Marc Odobez
IDIAP Research Institute
Martigny, Switzerland
odobez@idiap.ch

ABSTRACT

In the past, quantized local descriptors have been shown to be a good base for the representation of images, that can be applied to a wide range of tasks. However, current approaches typically consider only one level of quantization to create the final image representation. In this view they somehow restrict the image description to one level of visual detail. We propose to build image representations from multi-level quantization of local interest point descriptors, automatically extracted from the images. The use of this new multi-level representation will allow for the description of fine and coarse local image detail in one framework. To evaluate the performance of our approach we perform scene image classification using a 13-class data set. We show that the use of information from multiple quantization levels increases the classification performance, which suggests that the different granularity captured by the multi-level quantization produces a more discriminant image representation. Moreover, by using a multi-level approach, the time necessary to learn the quantization models can be reduced by learning the different models in parallel.

Categories and Subject Descriptors

I.4.10 [**Image Representation**]: Hierarchical; I.4.8 [**Scene Analysis**]: Object recognition—*scene classification*

General Terms

Vision, images, bag-of-visterms, bag-of-words, quantization, vocabularies, local descriptors, scene classification.

1. INTRODUCTION

Local image representation approaches are amongst the most sucessfull and versatile used in the field of computer vision Such approaches model the image content by subdividing the image into regions or parts on which individual features are computed. The resulting representation is then built as a collection of these local descriptors. As such, an alteration of an image part affects only some of the representation components, which render this approach particularly robust to partial occlusion, and allows for an accurate image content description.

Viewpoint invariant local descriptors [13, 10, 11] (i.e. features computed over automatically detected local interest areas) have proven to be useful in long-standing problems such as viewpoint-independent object recognition, wide baseline matching, and image retrieval. They provide robustness to image clutter, partial visibility, and occlusion. They were designed to have high degree of invariance, and, as a result, are robust to changes in viewpoint and lighting conditions.

Modelling images based on quantized local invariant features has also proven in recent years to provide a robust and versatile way to model images, providing good classification [9, 25, 17, 5, 1, 8], retrieval [19, 18] and segmentation of images [17, 4]. The great advantages of modelling images based on quantized local invariant features for the tasks of retrieval and classification are that the same methodology can be used for different image categories and that performance is normally better, or at worst similar, to most of the existing task specific state-of-the-art. This was shown in the case of images containing objects [25, 12, 9] and scenes [17, 5, 1].

One related area of research that has recently gained importance regards the investigation of methodologies for the quantization of the local descriptors [7, 12, 6]. The representation of images based on quantized local interest point descriptors is dependent on the quantization method on which it is based. When we perform quantization, we want to retain as much information as possible from the extracted local descriptors, so that our image description remains as faithful as possible. In current systems, K-means is the most widely used method, due mainly to its simplicity. However, new methods have been introduced that showed that better results can be obtained by using more efficient clustering methods [7, 12]. These methods create more discriminant vocabularies by using more complex quantization methodologies than a basic K-means.

In this paper, we present an approach to model scene images, using multi-level local interest descriptors quantization, a common idea in the image processing literature but which has not been explored yet for visterms vocabularies. This methodology remains simple, as it is still based on K-means clustering, but creates a more complete representation of the image due to the inclusion fine and coarse local image information in a joint approach. We will show that this new representation based on a multi-level quantization

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

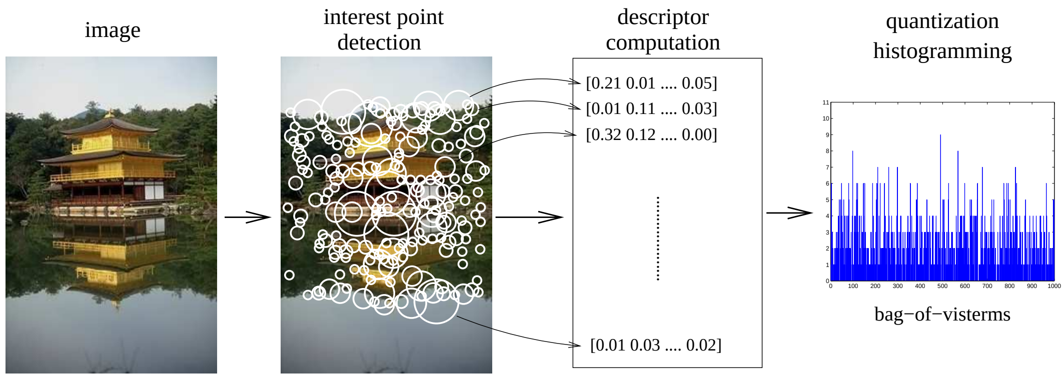


Figure 1: Computation steps of an image's BOV representation.

of the local interest point detectors increases performance without increasing complexity of our system. It must be said that although we, in this paper, refer only to the task of classification of images, the use of better quantization methods can result in a better performance in other tasks such as ranking/retrieval of images.

The rest of the paper is organized as follows. The next section discusses related work. Section 3 presents the image modeling approach, including our proposed multi-level representation. Section 4 describes the experimental setup. Classification results are provided and discussed in Section 4.4. Section 5 concludes the paper.

2. RELATED WORK

The problem of image modeling using low-level features has been studied in image and video retrieval for several years [22, 21, 15, 16, 20]. Broadly speaking, the existing methods differ by the definition of the target image classes, the specific image representations, and the classification method. We focus our discussion on the image representation.

Image representations based on quantized invariant local descriptors have been used for many tasks and with variations on both local detectors/descriptors and the subsequent image representation. Sivic et. al. [19] (but also other authors [3, 23]) proposed to cluster and quantize local invariant features into visterms, to increase speed and produce a more stable representation, for object matching in frames of a movie. However, quantizing local descriptors decreases the discriminative power, reducing the capacity to describe the local structure and texture. Nevertheless, such approaches allow to reduce noise sensitivity in matching and to search efficiently through a given video for frames containing the *same* visual content (e.g. an object) using inverted files.

The use of quantized local descriptors was further extended by Csurka et. al. [25] to the task of object recognition. Using the Caltech object image database the authors created a representation based on a histogram of the quantized local descriptors, inspired by the *bag-of-words* (BOW) representation in text. Despite the fact that no geometrical information is kept in the representation, this system was shown to outperform the state-of-the-art. More recently, Leibe et. al. [12] introduced a generative modeling approach for multiple object detection. This approach uses a hierarchical representation of visual object parts, in a generative framework. This allows the modeling of the parts' relationships together with the overall scale and orientation of the detected object, being able to model multiple object in each image.

In other work Bosch et. al. [1], Fei-Fei et. al. [5] and Quelhas et. al. [17], have show that the representation of images by quantized local descriptors can be further extended by using Latent Aspect Models. The results presented in the mentioned publications shown that good performance for scene image classification can be obtained by using quantized local descriptors to represent the image. The use of latent aspect modeling enabled experiments in clustering and ranking of images into meaningful groups.

Regarding the clustering methods used in the quantization process of previous work in this research area, K-means clustering is the most widely used approach. However some alternatives have been proposed. Jurie et. al. [7] proposed a new codebook construction methodology that increases performance by representing areas of the feature space with lower data density. The authors create the visual codebook using an acceptance-radius based clustering method. Their approach combines the advantages of on-line clustering and mean-shift, in an under-sampling framework. Mikolajczyk et. al. [12] used a hierarchical quantization methodology to create a hierarchical codebook to represent the local interest point descriptors. This clustering is then used for object recognition, in which parts of the object correspond to the clustering labels at different levels.

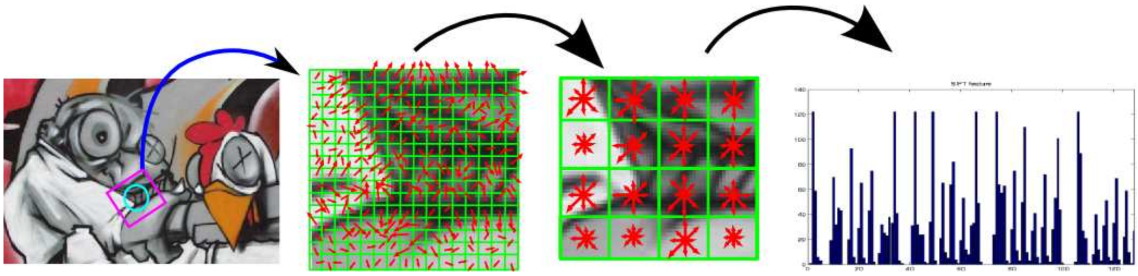
3. IMAGE MODELING

In this section we will start by describing the modeling of images based on quantized local interest point descriptors, the *bag-of-visterms* (BOV) -we will refer to quantized local interest point descriptors as *visterms* for the rest of this paper. We will then present the use of multi-level quantization which allows us to extend the standard BOV representation.

3.1 Bag-of-visterms image representation

The construction of the bag-of-visterms (BOV) feature vector h from an image I involves the different steps illustrated in Figure 1. In brief, interest points are automatically detected in the image, then local descriptors are computed over those regions. All the descriptors are quantized into visterms, and all occurrences of each specific visterm of the vocabulary in the image are counted to build the BOV representation of the image.

In the following we describe in more detail each step of



(a) (b) (c) (d)

Figure 2: Illustration of the SIFT feature extraction process. From left to right: (a) original image with the local interest point to describe, showing the detected location, scale and area used for sampling. (b) local interest area with gradient samples at each grid point, blue circle illustrates the Gaussian weighting window. (c) local individual orientation histograms which result of accumulating each sample into the corresponding bin of its local histogram. (d) final 128 dimensional SIFT feature (before normalization).

this methodology.

3.1.1 Local interest point detectors

The goal of interest point detectors is to automatically extract characteristic points -and more generally regions- from the image, which are invariant to some geometric and photometric transformations. This invariance property is interesting, as it ensures that given an image and its transformed version, the same image points will be extracted from both and hence, the same image representation will be obtained. Several interest point detectors exist in the literature. They vary mostly by the amount of invariance they theoretically ensure, the image property they exploit to achieve invariance, and the type of image structures they are designed to detect [10, 14, 11].

In this work, we use the difference of Gaussians (DOG) point detector [10]. This detector essentially identifies blob-like regions where a maximum or minimum of intensity occurs in the image, it is invariant to translation, scale, rotation and constant illumination variations. This detector was shown to perform well in comparisons previously published [14]. An additional reason to prefer this detector over fully affine-invariant ones [13, 11], is also motivated by the fact that an increase of the degree of invariance may remove information about the local image content that is valuable for classification.

3.1.2 Local interest point descriptors

Local descriptors are computed on the region around each interest point detected by the local interest point detector. There are several sources of information at the local level, as such local descriptors can be based on several different local properties. To describe the information around the detected local interest points we use the SIFT (Scale Invariant Feature Transform) local interest point descriptor [10]. The choice to use SIFT was motivated by the findings of several publications [19, 14, 5, 1], where SIFT was found to work best.

The SIFT descriptors is based on the gray-scale representation of the image. SIFT features are local histograms of edge directions computed over different parts of the interest region. They capture the structure of the local image regions, which correspond to specific geometric configurations of edges or to more texture-like content. In [10], it was shown that the use of 8 orientation directions and a grid of 4 × 4 parts gives a good compromise between descriptor size and accuracy of representation. The size of the feature vector is thus 128. Orientation invariance is achieved by estimating the dominant orientation of the local image patch and normalizing for rotation. Figure 2 illustrates the steps of the construction of the SIFT feature.

In this paper both the detection of the DOG local interest points and the computation of the SIFT local interest point descriptors are obtained using the implementations available at `http://lear.inrialpes.fr/people/dorko/`.

3.1.3 Vocabulary model construction

When applying the two preceding steps to a given image, we obtain a set of real-valued local interest point descriptors. We then quantize each local descriptor into one of a discrete set $\mathcal{V}$ of visterms v according to a nearest neighbor rule:

$$\begin{array}{l} s \longmapsto Q(s) = v_i \Longleftrightarrow \operatorname{dist}(s, v_i) \leq \operatorname{dist}(s, v_j) \\ \forall j \in \{1, \ldots, N_{\mathcal{V}}\} \end{array} \tag{1}$$

where $N_{\mathcal{V}}$ denotes the size of the visterm set. We will call vocabulary the set $\mathcal{V}$ of all visterms.

The vocabulary is constructed using K-means: the K-means algorithm is applied to a set of local descriptors extracted from training images, and the means are kept as visterms. We used the Euclidean distance in the clustering and choose the number of clusters depending on the desired vocabulary size. The choice of the Euclidean distance to compare SIFT features is common [10, 13, 5, 1].

We can inspect the resulting visterm vocabulary by viewing the local patches that were clustered together. Figure 3 shows some of the local image area patches that were clustered together by K-means. We can see that for a larger amount of clusters the resulting patches have a more similar appearance (Figure 3(bottom)). Which leads to a finer, more specific, image representation. Clusters from a model with less clusters tend to have a more noise appearance (Figure 3(top)). Which leads to a more generic image representation.

3.1.4 Bag-of-visterms construction

The standard BOV representation of the image is constructed from the local interest point descriptors according to:

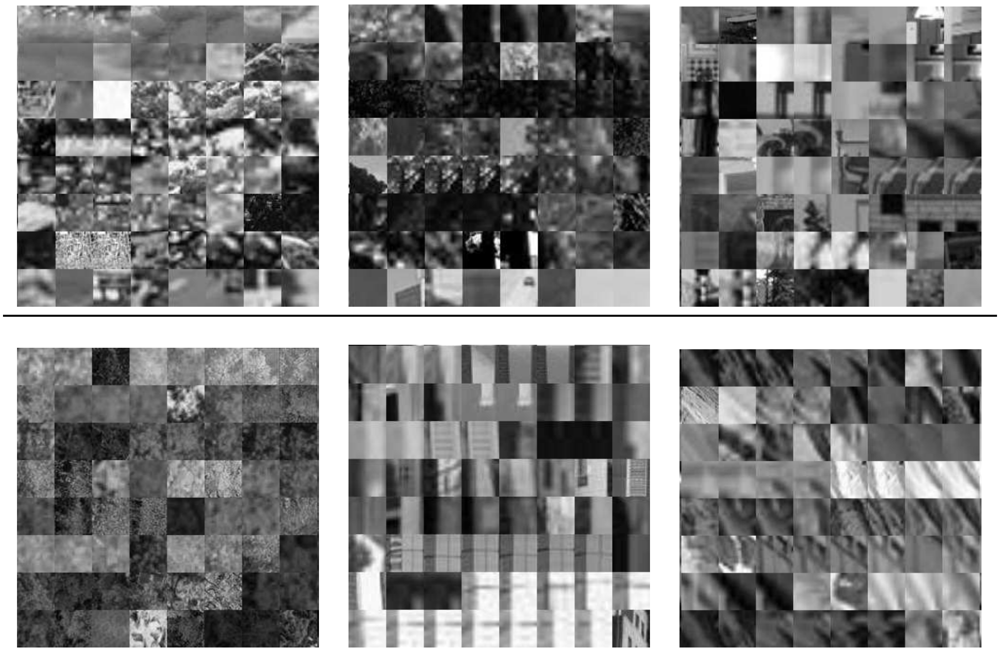


Figure 3: Image illustrating fine and coarse visterms. Each image is a group of local image patches that have been clustered together by K-means. Top: K-means with 100 clusters. Bottom: K-means with 3000 clusters.

$$h^{\mathcal{V}}(I) = [h_1^{\mathcal{V}}(I), ..., h_{N_{\mathcal{V}}}^{\mathcal{V}}(I)], \\ \text{with } h_i^{\mathcal{V}}(I) = \mathrm{n}(I, v_i, \mathcal{V}) \quad (2)$$

where $h(I, v_i, \mathcal{V})$ denotes the number of occurrences of visterm v_i from vocabulary $\mathcal{V}$ in image I.

This single-level BOV representation of the image describes each image using the occurrence information of each visterm in that image. This vector-space representation of an image contains no information about spatial relationship between visterms.

3.2 Multi-level bag-of-visterms

As an extension to the standard BOV construction methodology, we propose the use of several levels of visterm quantization to create a more thorough image representation. These multiple quantization levels are obtained by using several K-means models with different number of clusters.

The motivation for a multi-level representation comes from the way K-means describes the local descriptor's feature space. K-means clustering models the data by defining clusters at high density locations in the feature space. Given more clusters, we can obtain a finer representation of the feature space, capturing more specific visual content. On the other hand, using less clusters we obtain a more coarse space representation, capturing a more general overview of the image content (see Figure 3).

We expect the new representation to allow for better distance kernel computation between image representations. Let us consider the case of an image I and its BOV representation $h^{\mathcal{V}_f}(I)$ constructed from a fine vocabulary $\mathcal{V}_f$. If we add enough noise to the image and obtain the image I_{noisy}, it is possible that descriptors will be quantized with different visterms, and the representation $h^{\mathcal{V}_f}(I_{noisy})$ will start to differ from that of I. Since the representation does not keep information whether visterms are close or not in the feature space, the representation of I_{noisy} may differ from I as much as from I_d with a visual content unrelated to I.

In the case of a multi-level BOV representation of the image we want to retain both a fine level representation and a coarse level representation of the image. In this way we can expect our representation to retain the image content in a more thorough and faithful way.

The multi-level BOV representation of the image is constructed from the local descriptors according to the following method. Let us denote by $\mathcal{V}_a$ and $\mathcal{V}_b$ two different vocabularies. The new representation associated with these 2 vocabularies $\mathcal{V}_{a+b}$ is simply obtained by concatenating the BOV representation associated with vocabulary $\mathcal{V}_a$ and $\mathcal{V}_b$:

$$h^{\mathcal{V}_{a+b}}(I) = [h^{\mathcal{V}_a}(I), h^{\mathcal{V}_b}(I)], \quad (3)$$

The number of levels, however, is not limited to two, being extensible to as many as may be useful for a given task. In this approach, each local image feature contributes multiple times for the final image representation, one for each vocabulary level.

The final multi-level BOV representation is the concatenation of the several BOVs, each with an independent visterm vocabulary. In addition to providing different degrees of quantization granularity, the multi-level representation implicitly takes into account several K-means initialization (one per visterm vocabulary).

In the proposed representation we attribute the same relative importance to each level of the image's visterm description. It may be the case that for a specific scene class one vocabulary level may be more representative that another, motivating the introduction of a mixing parameter to weight the visterm vocabulary attribution before concatenation. However, this could affect the classification rates of other classes, and we did not follow this path.

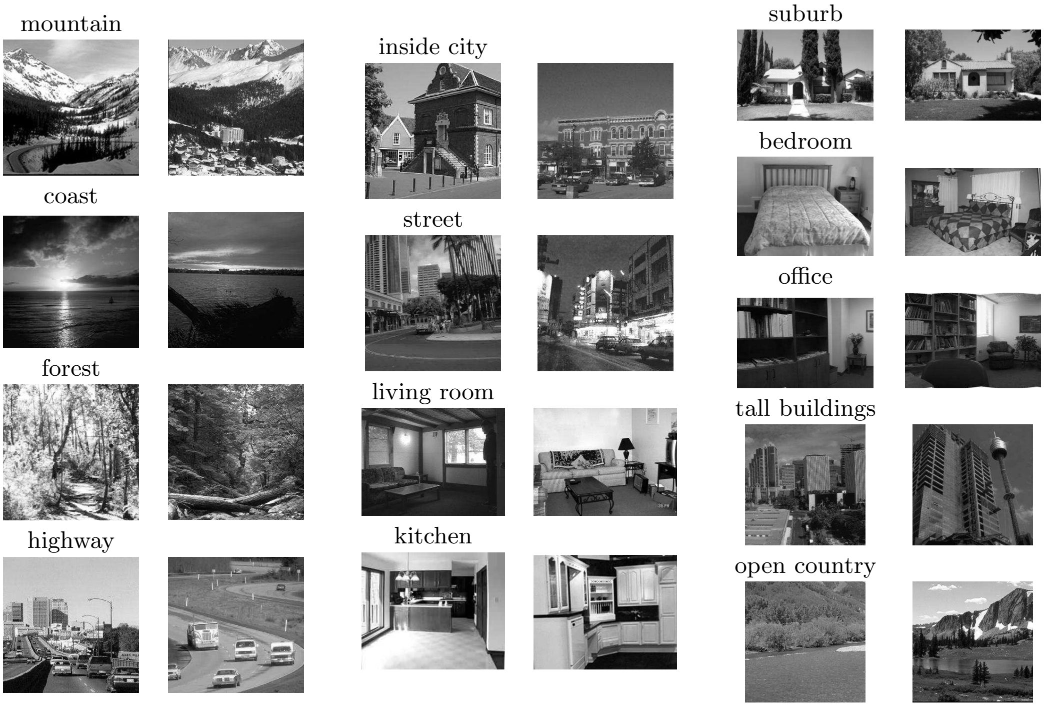


Figure 4: Images from the 13-class data set introduced by Fei-Fei et. al..

4. EXPERIMENTS AND DISCUSSION

In this section, we validate our proposed approach on the task of scene classification. First we present the data set used in the experiments, next we present the experimental setup and SVM setup. We then present the computational complexity for several vocabularies. Finally, we present the classification results for the different image representations.

4.1 Data set

In order to test our approach on the task of scene classification experiments we use the data set introduced by Fei-Fei et. al. [5]. In [5], the authors tackle the classification of 13 different scene types. This data set contains a total of 3859 images of approx. 60000 pixel resolution, varying in exact size and XY ratio. The images are distributed over 13 scene classes as follows (the number in parenthesis indicates the number of images in each class): bedroom (216), coast (360), forest (328), highway (260), inside city (308), kitchen (210), living room (289), mountain (374), open country (410), office (215), street (292), suburb (241), and tall buildings (356) [1]. See Figure 4 for some example images from the classes of this dataset.

This data set is challenging given the respective number of classes and the intrinsic ambiguities that arise from their definition. For example, images from the inside city and street categories share a very similar scene configuration.

4.2 Experimental setup

The protocol we followed for the scene classification experiments presented in the paper was the following. The full dataset of a given experiment was divided into 10 parts, thus defining 10 different splits of the full dataset. One split corresponds to keeping one part of the data for testing, while using the other nine parts for training.

The K-means models were created using 760 images randomly chosen from the training set from each split. Table 1 gives the computation time for the K-means model training for different k, using a Dual AMD Opteron 248 (2.190GHz, 1024Kb cache L2) machine with 4GB memory. As can be seen, the complexity grows with the size of the vocabulary, and can be quite large. Thus, since training 10 K-means models, one for each split, is time consuming we trained one model to be used by splits 1 through 5 and one for splits 6 through 10 (i.e., when testing on split 1 to 5, we used a vocabulary constructed from images of part 6 to 10). The remaining training data were used to cross-validate the SVM classifier.

We obtain 10 different classification results. The classification performance is obtained by averaging the performance of the classification results for each of the 13 classes as defined in [5]. The values we report for all experiments correspond to the average classification performance over all splits, and standard deviations of that performance are provided in parentheses after the mean value.

4.3 SVM classifier

To perform classification we employed Support Vector Machines (SVMs) [2]. We choose to use SVMs since they have proven to be successful in solving machine learning problems in computer vision and text categorization applications, especially those involving large dimensional input spaces. In the current work, we used Gaussian kernel SVMs, whose bandwidth was chosen based on a 4-fold cross-validation procedure, on the 4-folds of the training set that were not used

[1] http://faculty.ece.uiuc.edu/feifeili/data sets.html

Table 1: Computational cost in seconds of the model construction for the proposed modeling options.

Vocabulary	Training time (s)
V100	540
V300	1140
V600	1920
V1000	2820
V3000	13320
V5000	15600

Table 2: Scene image classification results using the 13-class data set introduced by Fei-Fei et. al.. The number in paranethesis represents te standard deviation of results across the splits. The $MLV1$ representation corresponds to the concatenation of the $V100$ and $V300$ BOV representations. $MLV2$ is built from $V100$, $V300$ and $V600$, $MLV3$ from $V100$ to $V1000$, and $MLV4$ from $V100$ to $V3000$.

Method	Classification Performance
Fei-Fei et. al.	65.2 (?)
BOV $V100$	66.2 (1.0)
BOV $V300$	68.1 (0.5)
BOV $V600$	68.2 (1.0)
BOV $V1000$	69.7 (1.1)
BOV $V3000$	69.6 (0.6)
BOV $V5000$	68.7 (1.2)
BOV $MLV1$	69.8 (1.2)
BOV $MLV2$	72.8 (1.0)
BOV $MLV3$	73.2 (0.6)
BOV $MLV4$	72.7 (1.3)

for vocabulary construction (see previous sub-section).

Standard SVMs are binary classifiers, which learn a decision function through margin optimization, such that function is large (and positive) for any input which belongs to the target class, and negative otherwise. To apply the SVM to our multi-class problem we adopt a one-against-all approach [24].

4.4 Classification results

First we show the results for the single-level BOV representations, then using the proposed multi-level representation, resulting from the concatenation of different vocabularies.

Table 2 shows the results of the single-level and multi-level representations for the 13-class scene classification problem. Even if a proper comparison is not possible due to the lack of information about variance, we compare our methods' performance with the original baseline from Fei-Fei et. al. [5].

We first consider the standard BOV approach, using vocabularies with sizes ranging from 100 to 5000 visterms ($V100$ to $V5000$). The standard BOV out-performs the baseline proposed by Fei-Fei et. al. [5] for most vocabulary sizes. The maximum performance is obtained for representations based on vocabularies of size 1000 and 3000 visterms, with a small degradation of performance for representations based on larger vocabularies.

As for our proposed approach we consider four multi-level representations $MLV1$, $MLV2$, $MLV3$ and $MLV4$. Where $MLV1$ results from the concatenation of the BOV representation obtained with vocabularies $V100$ and $V300$, $MLV2$ from the concatenation of $V100$, $V300$ and $V600$, $MLV3$ from the concatenation of $V100$, $V300$, $V600$, and $V1000$, and finaly $MLV4$ from the concatenation of $V100$, $V300$, $V600$, $V1000$ and $V3000$. Figure 6 illustrates on example of a $MLV4$ multi-level representation of an image, where we have highlighted the different vocabulary contributions. As we can see the use of a multi-level representations improves the classification results. This is even the case when the final dimension of the multi-level representation is the same as the vocabulary that gave the same results (i.e. $V1000$). The performance evolution of both the single-level and multi-level BOV representations with respect to the representation's dimensionality can be observed in Figure 5

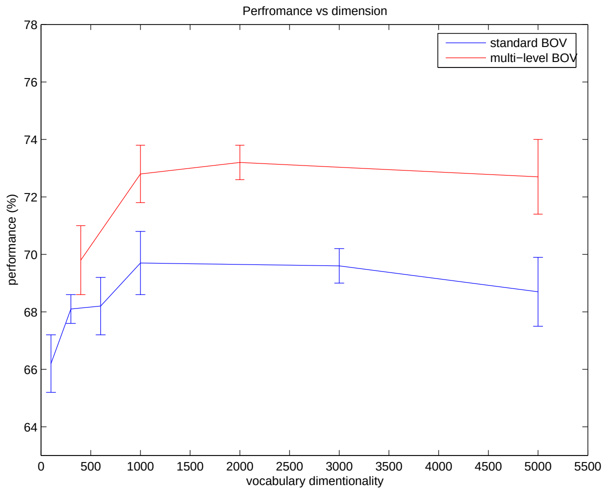


Figure 5: Graph of the performance versus dimensionality for each of the tested representations.

4.5 Discussion

There are several issues that might be worth discussing.

First, one may wonder whether the improvement is due to the fact that we have a multi-level representation, or due to the fact that these multi-level representation were built from vocabularies with different k-means initializations. Our experience is that k-means initialization only affects classification results marginally, especially for large vocabulary sizes or significant enough databases. Notice for instance that the variation of classification accuracy across splits is quite small (cf Table 2), despite the fact that for each vocabulary $Vxxx$, two vocabularies were learned (see experimental setup section). The improvement observed in our experiments is thus entirely due to the use of the multi-level representation.

A second concern is that the representation is redundant, and this is indeed true. Notice however that, since the vocabularies are built independently, all features that belong to a fine visterm (from a large vocabulary) do not necessarily belong to a single coarse visterm of smaller vocabularies. In other words, there is not an explicit hierarchy between visterms, and this makes the representation more powerful. Still, it would be very interesting to study the degree of redundancy that can be observed in practice, using for instant aspect models [17] that automatically identify visterms synonyms in an image collection through co-occurence analysis.

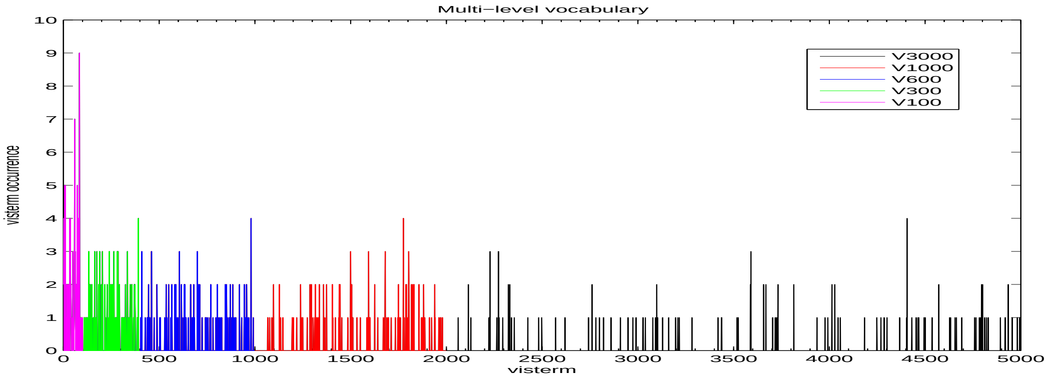


Figure 6: The multi-level BOV image representation $MLV3$. Each color represents one of the BOV representations that were concatenated to create the $MLV3$ multi-level representation.

5. CONCLUSION

Given the results we obtained in this paper, we have validated the hypotheses put forward in the introduction: that using information from different quantization levels can improve image representation and result in a better classification performance.

Based on the results presented in this paper, we believe that the presented extension to the standard bag-of-visterms representation is effective as a scene modeling methodology and adequate for scene classification tasks.

Although, in this paper, we present this new representation only in the context of a classification task, there may be other tasks for which the inclusion of a multi-level quantization information can be expected to bring similar improvements.

6. ACKNOWLEDGMENTS

This work was supported by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. The NCCR is managed by the Swiss National Science Foundation on behalf of the Federal Authorities. This work was also funded by the European project "CARTER: Classification of visual Scenes using Affine invariant Regions and TExt Retrieval methods" part of "PASCAL: Pattern Analysis, Statistical Modeling and Computational Learning ", through the Swiss Federal Office for Education and Science (OFES).

7. REFERENCES

[1] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via PLSA. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, Graz, Austria, May 2006.

[2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] A. de Vries. *Content and multimedia database management systems.* PhD thesis, Twente University, 1999.

[4] G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Nice, Oct. 2003.

[5] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *In Proceedings of IEEE Conference in Computer Vision and Pattern Recognition (CVPR)*, San Diego, Jun. 2005.

[6] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2005.

[7] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 604–610, 2005.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *In Proc. of IEEE CVPR*, 2006.

[9] B. Leibe, K. Mikolajczyk, and B. Schiele. Efficient clustering and matching for object class recognition. In *In Proceedings of British Machine Vision Conference (BMVC)*, 2006.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[11] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, Cardiff, Sep. 2002.

[12] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *Proceedings of IEEE CVPR*, New York, June 2006.

[13] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *In Proceeding of European Conference Computer Vision*, 2002.

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Toronto, June 2003.

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[16] S. Paek and C. S.-F. A knowledge engineering approach for image classification based on

probabilistic reasoning systems. In *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, Aug. 2000.

[17] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *In Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Beijing, Oct. 2005.

[18] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in image collections. In *Proceedings of IEEE International Conference on Computer Vision*, Beijing, October 2005.

[19] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of IEEE International Conference on Computer Vision*, Nice, October 2003.

[20] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[21] M. Szummer and R. Picard. Indoor-outdoor image classification. In *IEEE International CAIVD Workshop caivd (part of ICCV'98)*, Bombay, January 1998.

[22] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001.

[23] T. Westerveld and A. de Vries. Generative probabilistic models for multimedia retrieval: query generation versus document generation. *IEE Proceedings - Vision, Image and Signal Processing*, 152(6):852–858, 2005.

[24] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May 1998.

[25] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proceedings of LAVS Workshop, in ICPR'04*, Cambridge, August 2004.