Adapted Vocabularies for Generic Visual Categorization

Florent Perronnin, Christopher Dance, Gabriela Csurka, and Marco Bressan

Xerox Research Centre Europe, 6, chemin de Maupertuis, 38240 Meylan, France
{Firstname.Lastname}@xrce.xerox.com

Abstract. Several state-of-the-art Generic Visual Categorization (GVC) systems are built around a vocabulary of visual terms and characterize images with one histogram of visual word counts. We propose a novel and practical approach to GVC based on a universal vocabulary, which describes the content of all the considered classes of images, and class vocabularies obtained through the adaptation of the universal vocabulary using class-specific data. An image is characterized by a set of histograms - one per class - where each histogram describes whether the image content is best modeled by the universal vocabulary or the corresponding class vocabulary. It is shown experimentally on three very different databases that this novel representation outperforms those approaches which characterize an image with a single histogram.

1 Introduction

Generic Visual Categorization (GVC) is the pattern classification problem which consists in assigning one or multiple labels to an image based on its semantic content. We emphasize the use of the word "generic" as the goal is to classify a wide variety of objects and scenes. GVC is a very challenging task as one has to cope with variations in view, lighting and occlusion and with typical object and scene variations.

Several state-of-the-art GVC systems [14, 1, 4, 9, 16] were inspired by the *bag-of-words* (BOW) approach to text-categorization [13]. In the BOW representation, a text document is encoded as a histogram of the number of occurrences of each word. Similarly, one can characterize an image by a histogram of visual words count. The *visual vocabulary* provides a "mid-level" representation which helps to bridge the semantic gap between the low-level features extracted from an image and the high-level concepts to be categorized [1]. However, the main difference with text categorization is that there is no given visual vocabulary for the GVC problem and it has to be learned *automatically* from a training set.

To obtain the visual vocabulary, Sivic and Zisserman [14] and Csurka et al. [4] originally proposed to cluster the low-level features with the K-means algorithm, where each centroid corresponds to a visual word. To build a histogram, each feature vector is assigned to its closest centroid. Hsu and Chang [9] and Winn et al. [16] made use of the information bottleneck principle to obtain more discriminative vocabularies. Farquhar et al. also proposed a generative model, the Gaussian Mixture Model (GMM), to perform clustering [7]. In this case, a

A. Leonardis, H. Bischof, and A. Prinz (Eds.): ECCV 2006, Part IV, LNCS 3954, pp. 464–475, 2006.

low-level feature is not assigned to one visual word but to all words probabilistically, resulting in a continuous histogram representation. They also proposed to build the vocabulary by training class specific vocabularies and agglomerating them in a single vocabulary (see also the work of Leung and Malik [10] and Varma and Zisserman [15] for the related problem of texture classification). Although substantial improvements were obtained, we believe that this approach is unpractical for a large number of classes C. Indeed, if N is the size of the class-vocabularies, the size of the agglomerated vocabulary, and therefore of the histograms to be classified, will be $C \times N$ (c.f. the curse of dimensionality).

Our emphasis in this work is on developing a practical approach which scales with the number of classes. We define a *universal vocabulary*, which describes the visual content of all the considered classes, and *class vocabularies*, which are obtained through the *adaptation of the universal vocabulary* using class-specific data. While other approaches based on visual vocabularies characterize an image with a single histogram, in the proposed approach, an image is represented by a set of histograms of size $2 \times N$, one per class. Each histogram describes whether an image is more suitably modeled by the universal vocabulary or the corresponding adapted vocabulary.

The remainder of this paper is organized as follows. In section 2, we motivate the use of a universal vocabulary and of adapted class-vocabularies and describe the training of both types of vocabularies. In section 3, we show how to characterize an image by a set of histograms using these vocabularies. In section 4, we explain how to reduce significantly the computational cost of the proposed approach with a fast scoring procedure. In section 5, we show experimentally that the proposed representation outperforms those approaches which characterize an image with a single histogram. Finally, we draw conclusions.

2 Universal and Adapted Vocabularies

Let us first motivate the use of a universal vocabulary and of adapted class-vocabularies with a simple two-class problem where cats have to be distinguished from dogs.

A universal vocabulary is supposed to represent the content of all possible images and it is therefore trained with data from all classes under consideration. Since cats and dogs have many similarities, cats' and dogs' low-level feature vectors are likely to cluster into similar visual words such as "eye", "ear" or "tail". Hence, a histogram representation based on such a vocabulary is not powerful enough to help distinguish between cats and dogs. However, one can derive class vocabularies by adapting the universal vocabulary with class-specific data. Therefore, the universal "eye" word is likely to be specialized to "cat's eye" and "dog's eye" as depicted on Figure 1. Note that, although visual words are not guaranteed to be as meaningful as in the previous example, we believe that the combination of these universal and specific representations provides the necessary information to discriminate between classes.

As there exists a large body of work on the adaptation of GMMs, we represent a vocabulary of visual words by means of a GMM as done in [7]. Let us denote

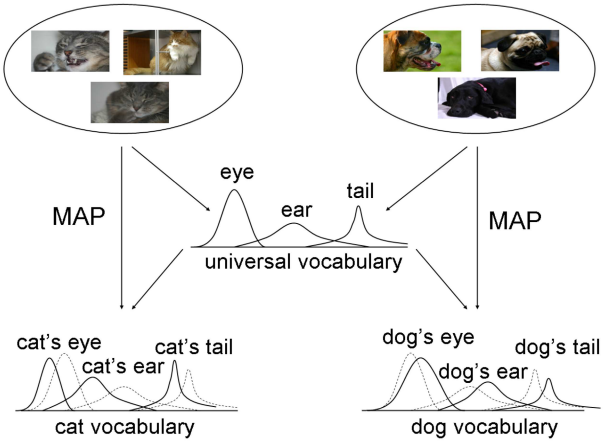


Fig. 1. The cats and dogs example: training a universal vocabulary with images from both classes and adapting this vocabulary to cat and dog vocabularies with class-specific data

by λ the set of parameters of a GMM. $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1...N\}$ where w_i, μ_i and Σ_i denote respectively the weight, mean vector and covariance matrix of Gaussian i and where N denotes the number of Gaussians. Each Gaussian represents a word of the visual vocabulary: w_i encodes the relative frequency of word i, μ_i the mean of the word and Σ_i the variation around the mean. In the following, we assume that the covariance matrices are diagonal as (i) any distribution can be approximated with an arbitrary precision by a weighted sum of Gaussians with diagonal covariances and (ii) the computational cost of diagonal covariances is much lower than the cost involved by full covariances. We use the notation $\sigma_i^2 = \text{diag}(\Sigma_i)$.

If an observation x has been generated by the GMM, we have:

$$p(x|\lambda) = \sum_{i=1}^{N} w_i p_i(x). \tag{1}$$

The components p_i are given by:

$$p_i(x) = \frac{\exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\}}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \tag{2}$$

where D is the dimensionality of the feature vectors and $|.|$ denotes the determinant operator.

We now explain how to train the universal and class vocabularies. The universal vocabulary is trained using maximum likelihood estimation (MLE) and the class vocabularies are adapted using the maximum a posteriori (MAP) criterion.

2.1 MLE Training of the Universal Vocabulary

Let $X = \{x_t, t = 1...T\}$ be the set of training samples. In the following, the superscript u denotes that a parameter or distribution relates to the universal

vocabulary. The estimation of λ^u may be performed by maximizing the log-likelihood function $\log p(X|\lambda^u)$. This is referred to as MLE.

The standard procedure for MLE is the Expectation Maximization (EM) algorithm [5]. EM alternates two steps: (i) an expectation (E) step where the posterior probabilities of mixture occupancy (also referred to as occupancy probabilities) are computed based on the current estimates of the parameters, and (ii) a maximization (M) step, where the parameters are updated based on the expected complete data log-likelihood which depends on the occupancy probabilities computed in the E-step.

For the E-step, one simply applies Bayes formula to obtain:

$$\gamma_t(i) = p(i|x_t, \lambda^u) = \frac{w_i^u p_i^u(x_t)}{\sum_{j=1}^N w_j^u p_j^u(x_t)}. \tag{3}$$

The occupancy probability $\gamma_t(i)$ is the probability for observation x_t to have been generated by the i-th Gaussian.

The M-step re-estimation equations are [2]:

$$\hat{w}_i^u = \frac{1}{T}\sum_{t=1}^T \gamma_t(i) \tag{4}$$

$$\hat{\mu}_i^u = \frac{\sum_{t=1}^T \gamma_t(i) x_t}{\sum_{t=1}^T \gamma_t(i)} \tag{5}$$

$$(\hat{\sigma}_i^u)^2 = \frac{\sum_{t=1}^T \gamma_t(i) x_t^2}{\sum_{t=1}^T \gamma_t(i)} - (\hat{\mu}_i^u)^2 \tag{6}$$

where x^2 is a shorthand for $\mathrm{diag}(xx')$.

Note that the initialization is an issue of paramount importance. Indeed EM is only guaranteed to converge to a local optimum and the quality of this optimum is largely dependent on the initial parameters. This initialization issue will be discussed in 5.

2.2 MAP Adaptation of Class Vocabularies

Let X be the set of adaptation samples. In the following, the superscript a denotes that a parameter or distribution relates to an adapted vocabulary.

The class vocabularies are estimated by adapting the universal vocabulary using the class training data and a form of Bayesian adaptation: MAP. The goal of MAP estimation is to maximize the posterior probability $p(\lambda^a|X)$ or equivalently $\log p(X|\lambda^a) + \log p(\lambda^a)$. Hence, the main difference with MLE lies in the assumption of an appropriate prior distribution of the parameters to be estimated. Therefore, it remains to (i) choose the prior distribution family and (ii) specify the parameters of the prior distribution.

The MAP adaptation of the GMM is a well-studied problem in the field of speech and speaker recognition [8, 12]. For both applications, one is interested in adapting a generic model, which reasonably describes the speech of any person,

to more specific conditions using the data of a particular person. It was shown in [8] that the prior densities for GMM parameters could be adequately represented as a product of Dirichlet and normal-Wishart densities. When adapting a generic model with MAP to more specific conditions, it is natural to use the parameters of the generic model as a priori information on the location of the adapted parameters in the parameter space.

As shown in [8], one can also apply the EM procedure to MAP estimation. During the E-step, the occupancy probabilities γ are computed as was the case for MLE:

$$\gamma_t(i) = p(i|x_t, \lambda^a). \tag{7}$$

The M-step re-estimation equations are [8]:

$$\hat{w}_i^a = \frac{\sum_{t=1}^T \gamma_t(i) + \tau_i^w}{T + \sum_{i=1}^N \tau_i^w} \ , \tag{8}$$

$$\hat{\mu}_i^a = \frac{\sum_{t=1}^T \gamma_t(i) x_t + \tau_i^m \mu_i^u}{\sum_{t=1}^T \gamma_t(i) + \tau_i^m} \ , \tag{9}$$

$$(\hat{\sigma}_i^a)^2 = \frac{\sum_{t=1}^T \gamma_t(i) x_t^2 + \tau_i^s \left((\sigma_i^u)^2 + (\mu_i^u)^2\right)}{\sum_{t=1}^T \gamma_t(i) + \tau_i^s} - (\hat{\mu}_i^a)^2 \ . \tag{10}$$

τ_i^w, τ_i^m and τ_i^s are relevance factors for the mixture weight, mean and variance parameters and keep a balance between the a priori information contained in the generic model and the new evidence brought by the class specific data. If a mixture component i was estimated with a small number of observations $\sum_{t=1}^T \gamma_t(i)$, then more emphasis is put on the a priori information. On the other hand, if it was estimated with a large number of observations, more emphasis will be put on the new evidence. Hence MAP provides a more robust estimate than MLE when little training data is available. The choice of parameter τ will be discussed in the section on experimental results.

3 Bipartite Histograms

Once the universal and adapted vocabularies have been properly estimated, we proceed as follows. For each class c, a novel vocabulary is obtained by merging the universal vocabulary and the adapted vocabulary of class c. This will be referred to as the *combined vocabulary* of class c. Note that the merging involves adjusting the weight parameters of the Gaussians to reflect the vocabulary size having doubled. In the case where the a priori probability $p(c)$ of class c is known, this can be done by multiplying the weights of the adapted vocabulary by $p(c)$ and the weights of the universal vocabulary by $(1 - p(c))$. The other parameters remain unchanged.

The rational behind this merging process is to make the Gaussians of the universal and adapted vocabularies "compete" to account for the feature vectors of an image. Indeed, if an image belongs to class c, it is more suitably described by the visual words of class c rather than by the words of the universal vocabulary.

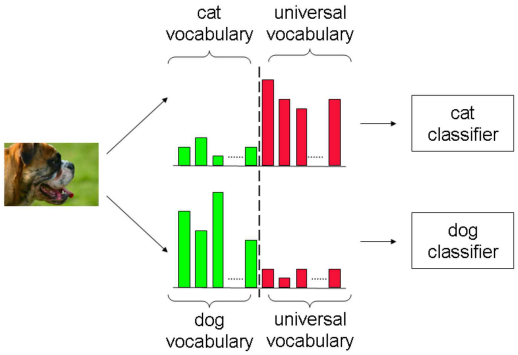


Fig. 2. Generating one bipartite histogram per category. Each histogram is subsequently fed to a different classifier.

On the other hand, if an image belongs to another class, then the visual words of the universal vocabulary will describe it more appropriately.

An image can therefore be characterized by a set of histograms - one per class - using these combined vocabularies. These histograms are said to be bipartite as half of the histogram reflects the contribution of the universal vocabulary in explaining the image while the other half reflects the contribution of the adapted vocabulary (c.f. Figure 2).

Interestingly, for a given image, summing the two halves of the bipartite histograms (i.e. summing the count of a word in the universal vocabulary part with the count of the corresponding word in the adapted vocabulary part) should lead to the same histogram approximately, whatever the class. Note that this histogram is the one we would obtain using only the universal vocabulary representation. Hence, the key of the success of the proposed approach is the ability to *separate for each class the relevant information from the irrelevant information.*

To classify these histograms, we use one Support Vector Machine (SVM) classifier per class. Each SVM is trained in a one-vs-all manner as done in [1, 4]. However, in [1, 4], as images are characterized by a single histogram, the same histograms are fed to the classifiers. In the proposed approach, each classifier is fed with different histograms, both at training and test time. Going back to our cats and dogs example, a "cat" classifier will be trained with histograms computed on the combined vocabulary of the class cat. In the same manner, at test time the histogram obtained with the combined vocabulary of the class cat will be fed to the cat classifier and the histogram obtained with the combined vocabulary of the class dog will be fed to the dog classifier.

4 Computational Cost

When estimating a histogram, the most intensive part is the Gaussian computation, i.e. the computation of the values $p_i(x)$ (c.f. equation (2)). If N is the

number of Gaussians in the universal vocabulary, and C is the number of classes, a direct implementation would require $N \times (C + 1)$ Gaussian computations per image. This is unacceptable for large values of N or C.

To reduce the computational cost, we make use of a fast scoring procedure devised by Reynolds et al. for the speaker recognition problem [12]. This technique is based on two observations. The first one is that, when a large GMM is evaluated, only a few of the mixtures will contribute significantly to the likelihood value (c.f. equation (1)) and therefore, only a few of the mixtures will have a significant occupancy probability $\gamma_t(i)$. This property was observed empirically. The second one is that the Gaussians of an adapted vocabulary retain a correspondence with the mixtures of the universal vocabulary. Therefore, if a feature vector x scores high on the i-th component of the universal vocabulary, it will score highly on the i-th Gaussians of all adapted vocabularies.

The fast scoring procedure operates as follows on each feature vector x_t:

1. Compute the likelihood $p_i^u(x_t)$ for all the mixture components i of the universal vocabulary (N Gaussian computations). Retain the K best components.
2. Compute the likelihood values $p_i^a(x_t)$ for the K corresponding components of the C adapted vocabularies ($K \times C$ Gaussian computations).
3. For the C combined vocabularies, compute the occupancy probabilities $\gamma_t(i)$ on the $2 \times K$ corresponding components. Assume that the occupancy probabilities are zero for the other components.

Hence, the number of Gaussian computations is reduced from $N \times (C + 1)$ to $N + K \times C$. For large values of C this is reduction of the computational cost by a factor N/K. Typical values for N and K are $N = 1,024$ and $K = 5$. Note that we did not observe any significant decrease of the performance in our experiments with as little as $K = 2$ best components. Hence the value $K = 5$ is a rather conservative choice.

Returning to our cats and dogs example, this fast scoring procedure simply consists in first determining whether the input feature vector corresponds to an eye, a tail, etc. and then if it is a tail, whether it is more likely to be the tail of cat or the tail of a dog.

5 Experimental Validation

In this section, we carry out a comparative evaluation of the proposed approach on three very different databases: an in-house database of scenes, the LAVA7 database and the Wang database. The two approaches which will serve as a baseline are (i) the one which makes use only of the universal vocabulary (as in [14, 4]) and (ii) the one which agglomerates class-vocabularies into a single vocabulary (as in [7]). We consider a classification task, i.e. each image is to be assigned to one class and the measure of performance is the percentage of images assigned to their correct classes. In the following section, we describe the experimental setup. We then provide results.

5.1 Experimental Setup

The low-level local features are based on local histograms of orientations as described in [11]. These features were extracted on a regular grid at different scales. As all images were resized before the feature extraction step so that they contained (approximately) the same number of pixels, the same number of features was extracted from all images (approximately).

The dimensionality of feature vectors was subsequently reduced from 128 to 50 using Principal Component Analysis (PCA). This decorrelates the dimensions of the feature vectors and thus makes the diagonal covariance assumption more reasonable. Discarding the last components also removes noise and thus increases the performance. It also significantly reduces the cost of Gaussian computations.

To alleviate the difficult initialization problem when training the universal vocabulary with MLE, we used a strategy inspired by the vector quantization algorithm. We start with a vocabulary of one unique word and then increase the number of Gaussians iteratively. Each iteration consists of two steps: (i) all the Gaussians which were estimated at the previous step with more than a given number of observations are split into two by introducing a slight perturbation in the mean and (ii) EM is performed until convergence, i.e. until the log-likelihood difference between two iterations falls below a predefined threshold. These two steps can be repeated until the desired number of Gaussians is obtained. An advantage of increasing progressively the number of Gaussians is that it allows to monitor the recognition performance to select the optimum vocabulary size.

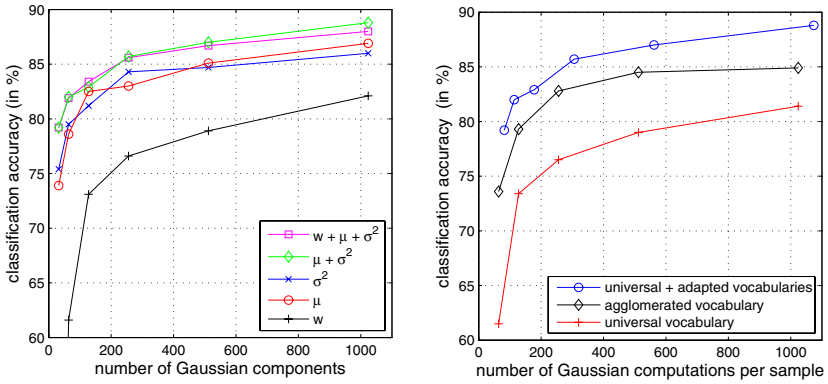
For MAP adaptation, to reduce the number of parameters to hand-tune, we enforced $\tau_i^w = \tau_i^m = \tau_i^s = \tau$. We tried different values for τ and found that values between 5 and 50 were reasonable. In our experiments, we set $\tau = 10$. We demonstrate below the influence of adapting either all parameters, i.e. the mixture weights, means and covariances, or a subset of the parameters.

As for classifying the histograms, we used linear SVMs for both the proposed approach and the approach based on a single vocabulary. The only parameter to set is the one which controls the trade-off between the margin and the number of misclassified points, commonly known as C. It was fixed to 300 in all the following experiments. Note that in the linear case the cost of classifying a histogram is independent of the number of support vectors and can be neglected compared to the cost of Gaussian computations.

5.2 Results

In-house database. The first set of experiments was carried out on an in-house database of 8 scenes relating to amusement parks, boats, New York city, tennis, sunrise/sunset, surfing, underwater and waterfalls. This is a challenging set as we collected the training data while the test material was collected independently by a third party. Approximately 12,000 images were available for training and 1,750 for testing.

We first determine which Gaussian parameters are the most crucial ones to adapt in the proposed approach. Results are presented on Figure 3(a) as the



(a) (b)

Fig. 3. Results on the in-house 8 scenes database. (a) Influence of the adaptation of the different Gaussian parameters (weight w, mean μ and covariance σ^2) on the classification accuracy. (b) Comparison of the proposed approach (universal + adapted vocabularies) with the two baseline systems (universal vocabulary and agglomerated vocabulary).

classification accuracy versus the number of Gaussian components, i.e. the vocabulary size. Clearly, adapting only the weights leads to a poor performance. Adapting either the means or the covariances has roughly the same impact and adapting both parameters leads to an additional small improvement. However, adapting the three parameters does not give further improvement. This experiment clearly shows that the relative frequency of a word (weight) in an adapted vocabulary has little influence; what matters is the location of the word (mean) and its variations (covariance). In the following, we adapt only the means and covariances.

We now compare the proposed approach with the two baseline approaches. Results are presented on Figure 3(b) as the classification accuracy versus the number of Gaussian computations per sample. For the two baselines, the number of Gaussian computations per sample is exactly the number of components. For the proposed approach, this is slightly higher (c.f. section 4). The proposed approach clearly outperforms the baselines. Indeed, it achieves an 88.8% accuracy while the approach based solely on a universal vocabulary achieves 81.4% accuracy and the approach based on an agglomerated vocabulary achieves an 84.9% accuracy for a vocabulary size of 1,024 visual words. This shows that the adapted vocabularies encode more discriminative information.

LAVA7 Database [4]. This database, also sometimes referred to as Xerox7 database [17], contains 1,776 images of seven objects: bikes, books, buildings, cars, faces, phones and trees. It served as a testbed for object recognition experiments during the course of the European LAVA project. The standard setup for running experiments on this database is a ten-fold cross-validation. Results are presented on Figure 4(a) as the classification accuracy versus the number

of Gaussian computations per sample. We can see that the proposed approach outperforms the two baseline systems.

To the best of our knowledge, the best results reported on this database are those of Zhang et al. [17]. With their approach, which makes use of two feature extractors, two feature descriptors and an earth mover's distance (EMD) based kernel, they achieve a 94.3% accuracy but at a very high computational cost: the classification of an image takes on the order of 1 min on a modern PC. Running our non-optimized code on a 2.4 GHz AMD Opteron with 4GB RAM, our best system categorizes an image into one of the 7 categories with an accuracy of 95.8% in roughly 150 ms: approximately 125ms for the feature extraction and 25ms for the histogram building (the cost of the SVM classification can be neglected).

Wang Database [3]. This database contains 10 categories: Africa, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountains and food. Each category contains 100 images, which makes a total of 1,000 images. We used the same setup as in [3]: we randomly divided each category set into a training set and a test set, each with 50 images, and repeated the experiment 5 times. To prove that our good results are not restricted to SIFT-like features, we experimented with color features based on local mean and standard deviation in the RGB channels. Results are presented on Figure 4(b) as the classification accuracy versus the number of Gaussian computations per sample. We can observe that the proposed approach performs best, thus proving that our good results are not SIFT-specific. If we run separately two systems, one based on SIFT features and one based on color features, and if we do a late fusion (averaging the scores of the two systems), we get a 92.8% classification accuracy. To the best of our knowledge, the highest accuracy which had been previously reported on this database was 87.3% [6].

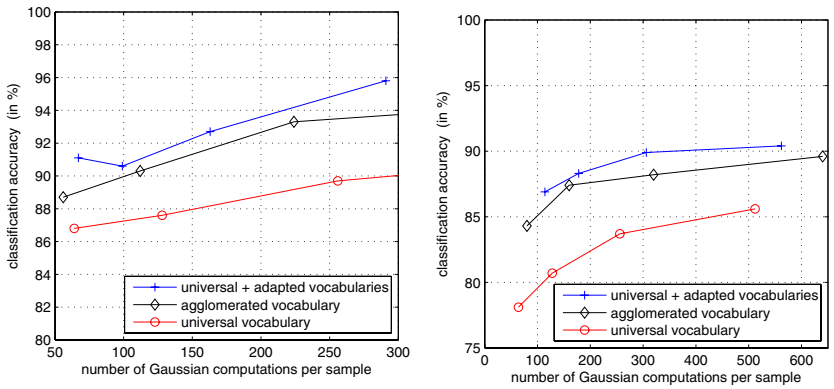


Fig. 4. Comparison of the proposed approach (universal + adapted vocabularies) with the two baseline systems (universal vocabulary and agglomerated vocabulary) on (a) the LAVA7 database and (b) the Wang database

6 Conclusion

We proposed a novel and practical approach to GVC based on a universal vocabulary, which describes the content of all the considered classes of images, and class vocabularies obtained from the universal vocabulary using class-specific data and MAP adaptation. An image is characterized by a set of histograms - one per class - where each histogram describes whether the image content is best modeled by the universal vocabulary or the corresponding class vocabulary. It was shown experimentally on three very different databases that this novel representation outperforms those approaches which characterize an image with a single histogram.

Note that, although less emphasis has been put on the reduction of the memory requirements, a simple approach could be used, if necessary, to reduce the number of Gaussians to store for each adapted vocabulary. As there exists a correspondence between the Gaussians in the universal and adapted vocabularies, one could save only those Gaussians which have significantly changed in the adapted vocabularies. This can be measured using various metrics such as the divergence, the Bhattacharya distance or the Gaussian overlap.

Also, although we have only considered a flat hierarchy of classes in this work, the proposed framework would be particularly suited to a hierarchical organization where the vocabularies of classes at a given level of the hierarchy would be adapted from their parent vocabularies.

Acknowledgments

This work was partially supported by the European project IST-2001-34405 LAVA (http://www.l-a-v-a.org) and the European project FP6-IST-511689 RevealThis (http://sifnos.ilsp.gr/RevealThis).

References

1. A. Amir, J. Argillander, M. Berg, S.-F. Chang, M. Franz, W. Hsu, G. Iyengar, J. Kender, L. Kennedy, C.-Y. Lin, M. Naphade, A. Natsev, J. Smith, J. Tesic, G. Wu, R. Yang, and D. Zhang. IBM research TRECVID-2004 video retrieval system. In *Proc. of TREC Video Retrieval Evaluation*, 2004.
2. J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical report, Department of Electrical Engineering and Computer Science, U.C. Berkeley, 1998.
3. Y. Chen, and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Mearning Research*, 5(2004):913–939, 2004.
4. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. of ECCV Workshop on Statistical Learning for Computer Vision*, 2004.
5. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
6. T. Deselaers, D. Keysers, and H. Ney. Classification error rate for quantitative evaluation of content-based image retrieval systems. In *Proc. of ICPR*, 2004.

7. J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. Technical report, University of Southampton, 2005.
8. J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, Apr 1994.
9. W. H. Hsu and S.-F. Chang. Visual cue cluster construction via information bottleneck principle and kernel density estimation. In *Proc. of CIVR*, 2005.
10. T. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. In *Proc. of ICCV*, 1999.
11. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.
12. D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.
13. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
14. J. S. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*, volume 2, pages 1470–1477, 2003.
15. M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *Int. Journal of Computer Vision*, 62(1–2):61–81, 2005.
16. K. Winn, A. Criminisi, and T. Minka. Object categorization by learned visual dictionary. In *Proc. of ICCV*, 2005.
17. J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: an in-depth study. INRIA, Research report 5737, 2005.