Marginal-based Visual Alphabets for Local Image Descriptors Aggregation*

Miriam Redi
EURECOM, Sophia Antipolis
2229 route des crêtes
Sophia-Antipolis
redi@eurecom.fr

Bernard Merialdo
EURECOM, Sophia Antipolis
2229 route des crêtes
Sophia-Antipolis
merialdo@eurecom.fr

ABSTRACT

Bag of Words (BOW) models are nowadays one of the most effective methods for visual categorization. They use visual dictionaries to aggregate the set of local descriptors extracted from a given image. Despite their high discriminative ability, one of the major drawbacks of BOW still remains the computational cost of the visual dictionary, built by clustering in the high dimensional feature space.
In this paper we introduce a fast, effective method for local image descriptors aggregation that is based on marginal approximations, i.e. the approximation of each descriptor component distribution. We quantize each dimension of the feature space, obtaining a visual *alphabet* that we use to map the image descriptors in a fixed-length visual signature. Experimental results show that our new method outperforms the traditional BOW model in both accuracy and efficiency for the scene recognition task. Moreover, we discover that the marginal-based aggregation provides complementary information with respect to BOW, by combining the two models in a video retrieval system based on TRECVID 2010 [9].

Categories and Subject Descriptors

I.4.7 [**Artificial Intelligence**]: Scene Analysis

General Terms

Algorithms

Keywords

Scene Recognition, Feature Extraction, CBIR

1. INTRODUCTION

Effective techniques for automatic image categorization are essential to manage large collections of digital images and video. The general approach is to model the redundant image information with a low-dimensional description, namely an image signature, and then classify such signatures with supervised learning techniques. Visual descriptions based on quantized local invariant features (e.g. [4]) have been extensively used for image categorization and retrieval. Among these approaches, the Bag-Of-Visual-Words [1] model has been proven to be a very successful method to aggregate discriminative local image properties. The image is first represented with a set of independent local patches encoded by robust local descriptors. The resulting representation is a high-dimensional, variable length description of the image. Visual dictionaries are then used to map each image to a fixed length feature vector that can be used as input for traditional classifiers (e.g. Support Vector Machines).
Many methods have been proposed to generate such dictionaries. Generally, they use clustering techniques on the local descriptors of a set of training images to define visual words. K-means [1] clustering is nowadays the most common technique for creating BOW, but recently more efficient methods based on mean-shift [2], hierarchical clustering [5], or fixed quantization based on lattices [10] have been proposed to improve the model performances.
Despite their good performances in image recognition and retrieval, one of the major drawbacks of all these approaches is their computational cost, for both clustering in the high dimensional feature space (determined by the descriptors distribution) and visual words assignment.
In this paper we present a simple, fast and effective algorithm for local feature quantization that we name MEDA (Marginal Estimation for Descriptors Aggregation). This approach provides a different way of aggregating local descriptors that does not involve any clustering or operation in the high-dimensional space, leading to an image signature that requires much less computation and provides better accuracy compared to traditional BOW models.

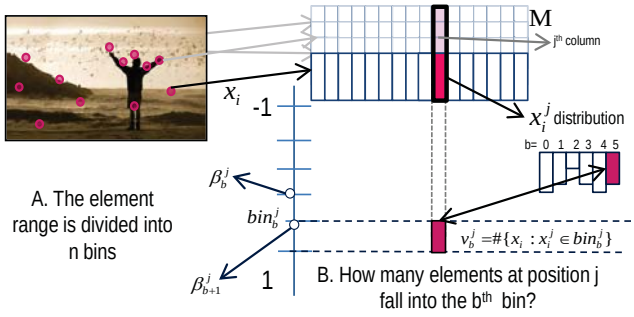

Figure 1: MEDA: a histogram representing the component distribution over unidimensional bins.

*Area chair: Nicu Sebe

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

Similar to the BOW model, a k-dimensional local invariant descriptor (LID) is used to describe a set of interest points in the image. While generally the quantization is performed in a k-dimensional space determined by the LID length, the basic idea of our approach is to model the k-dimensional space by the k marginal distributions approximations. This is obtained with an aggregation process that involves (see Fig. 1 for a visual explanation) two steps: (a) we quantize the range of each dimension of the LID into n bins, defining a reduced set of possible values that each of the k components of a described point can take; (b) given an image and its set of descriptors, we count the frequencies of the computed bin values, and we collect them in a $k \times n$ histogram, i.e. the MEDA image signature. Therefore, the marginal, i.e. the probability distribution of each component of the LID is approximated by a histogram representing its frequency over unidimensional bins.

Following the textual metaphor of the BOW, our method defines, for each component of the LID, a set of possible 1-d visual *letters*, namely the bin values; the collection of such *letters* is a *visual alphabet* that allows the mapping of an image to a fixed-length attribute vector. In this paper, we present and compare three different methods that define the values in the visual *alphabet*, based on different types of range quantization, namely uniform quantization, quantile-based quantization and an entropy-based quantization we perform using a decision tree.

In order to test the effectiveness of MEDA for image representation we first choose the indoor (67 classes) [7] and outdoor (8 classes) [6] scene recognition task. Results show that, in both databases, our technique outperforms in accuracy and efficiency the classical BOW model for the same feature size. We also evaluate the performances of our technique for the Trecvid 2010 [9] Semantic Indexing Task. We show that BOW and MEDA model achieve the same results for concept detection. Moreover, when combining MEDA with the BOW, we also discover that our new technique brings a new, complementary source of discriminative information in the local image analysis, improving the final retrieval precision by 25 %.

The remainder of this paper is organized as follows: Sec. 2 presents a detailed implementation of the MEDA model for local features quantization; we then explain, in Sec. 3, a variety of methods to approximate the range of the components; finally, in Sec. 4 we evaluate the performances of our model comparing it with traditional BOW.

2. MARGINALS ESTIMATION FOR DESCRIPTORS AGGREGATION

We propose an image representation that collects in a histogram the frequency of each component of the locally extracted vectors. While the BOW model quantizes the local features in a multi-dimensional space (*words*) determined by the descriptor length, here the quantization is performed in a 1-d space, for each component (*letter*) of the LID.

Given an image I, a set of w salient points is automatically detected in the image. Then, local descriptors $x_1, x_2, \ldots, x_w$ of length k are computed over the surrounding regions; we therefore obtain a set of normalized vectors $x_i = (x_i^1, \ldots, x_i^k)$, where each element x_i^j represents the value of the descriptor x_i at position j, $j = 1, \ldots, k$.

After normalization, each element x_i^j can take a value in the finite interval $R = [-1, 1]$, which covers a very large set of possible discrete values $a_1, a_2, \ldots, a_m$. The idea here is to quantize R by mapping it into a smaller set of discrete values $\beta^j \in R$, corresponding to a set of bins $bin_1^j, \ldots, bin_n^j$, i.e. our *alphabet*, defined for each dimension of the LID:

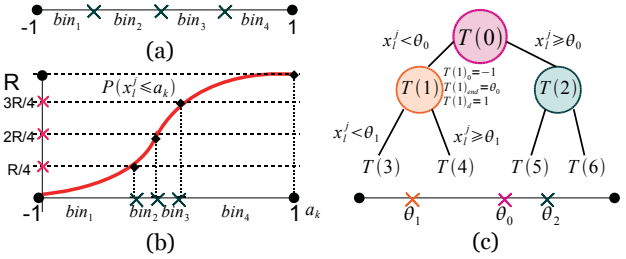

Figure 2: Three versions of the MEDA alphabet

$$bin_b^j = [\beta_b^j, \beta_{b+1}^j[, b = 0, \ldots, n-1, n < m \quad (1)$$

(+1 is added to the last bin). By doing so, each element in the image can be represented by the index of corresponding bin $bin_b^j : \beta_b^j \leq x_i^j < \beta_{b+1}^j$. The choice of the bin boundaries values will be discussed in the next Section.

We have therefore defined a set of shared visual *letters* that can be used to approximate the marginal distribution of the j^{th} element of the descriptors in the image. We can now represent the image as the collection of the number of elements x_i^j, $\forall i, j$ that fall into each of the identified bins.

The resulting signature for the image I is a vector

$$v = (v_1^1, v_2^1, \ldots, v_n^1, v_1^2, \ldots, v_n^k)$$

with $v_b^j = \#\{x_i : x_i^j \in bin_b^j\}$, $\forall i, j$,[1] The dimension of the MEDA signature is therefore $n \times k$.[2]

3. ALPHABET CONSTRUCTION

How to define the boundaries of such bins, our *letters*, so that the marginal of the j^{th} component, $\forall j$ is properly estimated? In this section we tackle this issue using three different approaches, namely: (1) **uniform quantization**: the range is divided into n equally spaced bins (2) **quantile-based quantization**: the range is divided so that the probability of a sample to fall into a bin is equal for all the n bins in the quantized space (3) **tree-based quantization**: for each bin, the boundaries are learnt by minimizing the overall entropy given a progressively smaller interval of R. Each of these methods leads to a different version of the MEDA histogram, that will be evaluated in Sec 4.

3.1 Uniform Bins

The most simple approach to define the bin boundaries over the data range is the uniform quantization. The advantage of such a simple approach is that it does not require prior knowledge of the marginal distribution of the components. As every component of the LID can take values in the same interval [-1,1], the resulting *alphabet* is an identical set of *letters* for every j. The range R is divided into n equal intervals of length $2/n$, and the set of bins valid for every component (see Fig. 2(a)) is defined as:

[1] $\#\{\cdot\}$ is a function that counts the number of the elements that satisfy the condition in brackets.

[2] Intuitively, if we define a matrix M whose rows are the descriptors x_i, v represents the concatenation of the n-dimensional histograms of the set of points stored in each column of a M

$$bin_b = [-1 + \tfrac{2b}{n}, -1 + \tfrac{2(b+1)}{n}[$$

3.2 Quantile-Based Bins

Here we try to adapt the width of each bin to the probability distribution of the component over the data range. This process requires a learning phase in which we identify the probability of the j^{th} component of the descriptors to take the value a_r, $r = 1, \ldots, m$ in the range R, i.e. the marginal distribution of x_i^j (see Fig. 2(b)). We need a dataset of N images over which we collect W described points x_l, $l = 1, \ldots, W$; we can then define the marginal:

$$p(a_r^j) = \#\{x_l : x_l^j = a_r\}$$

and the cumulative probability $P(x_l^j \leq a_r) = \sum_{s=1}^{r} p(a_s^j)$. We want each component x_l^j to be equally probable for all the bins in the range: we need therefore to find those values in the interval for which $p(x_l^j \in bin_b^j) = W/n$ for all b, being $\sum p(a_r^j) = W$. The final set of bins is defined as:

$$bin_b^j = [a_r^j : P(x_l^j \leq a_r) = \tfrac{bW}{n}, a_r^j : P(x_l^j \leq a_r) = \tfrac{(b+1)W}{n}[$$

3.3 Entropy-Based Bins

We propose a partition of the data range into a set of unbalanced bins, selected based on the minimization of the overall entropy. Here again we need a learning phase on a training set of N images and a total of W keypoints x_l. We build, for each position j of the LID, a decision tree T^j, with n splits built in n iterations, that progressively learns the boundaries of each bin_b^j.

Each node $T^j(t)$, at depth $T_d^j(t)$ of the tree considers the set of x_l^j that take values between $T_0^j(t)$ and $T_{end}^j(t)$. The tree growing starts from the root node $T^j(0)$, corresponding to the whole set of $x_l^j \in R$ and, at each step, finds the value θ_t^j in R for which the resulting partition of the data has the minimum entropy, i.e. the optimum bin boundary.

If we assume the dataset is categorized in c classes $y_1, .., y_c$, the general entropy of the data for a split $a_r \in R$ is:

$$H(y|a_k) = -p(x_l^j < a_r) \sum_{p=1}^{c} p(y_p|x_l^j < a_r) log(p(y_p|x_l^j < a_r))$$
$$-p(x_l^j \geq a_r) \sum_{p=1}^{c} p(y_p|x_l^j \geq a_r) log(p(y_p|x_l^j \geq a_r))$$

with $p(y_p|x_l^j < a_r)$ and $p(y_p|x_l^j \geq a_r)$ being the probability of a component belonging to an image labeled with category y_p to fall into the low/high bin generated by the split.[3] The following is the pseudo-code that summarizes how to grow a decision tree to learn the *alphabet* for the j^{th} component:

Grow_Tree
  $T^j(0) = \{root\}$
  repeat
    choose unmarked leaf $T^j(t)$
    find $\theta_t^j = \arg\min_{a_r} H_t(y|a_r)$, $T_0^j(t) \leq x_l^j < T_e^j nd(t)$
    if $T(t)_d^j < max\_depth$ **then**
      $T_0^j(t+1) \leftarrow T_0^j(t)$, $T_{end}^j(t+1) \leftarrow \theta_t^j$ {left child}
      $T_0^j(t+2) \leftarrow \theta_t^j$, $T_{end}^j(t+2) \leftarrow T_{end}^j(t)$ {right child}
    else
      mark $T^j(t)$
    end if
  until all leaves are marked
  $\beta_b^j \leftarrow$ in order tree walk on θ_t^j

[3] $p(y_p|x_l^j < a_r) = \frac{\#\{x_l^j : x_l^j < a \in y_p\}}{\#\{x_l^j : x_l^j < a_r\}}$; $p(y_p|x_l^j \geq a_r) = \frac{\#\{x_l^j : x_l^j \geq a \in y_p\}}{\#\{x_l^j : x_l^j \geq a_r\}}$

Two child nodes $T^j(1)$ and $T^j(2)$ are created as the result of the split at the first iteration (see Fig. 2 (c)); at the second iteration, $T^j(1)$ will find the best split for the set of elements for which holds $-1 \leq x_l^j < \theta_0$, while $T^j(2)$ will consider those x_l^j that lie between θ_0^j and 1. The process is iterated until the maximum depth (max_depth) required to identify n bins is reached. Finally, the set of boundaries θ_t^j found is sorted and the bin values are assigned according to Eq. (1).

4. EXPERIMENTAL VALIDATION

This section presents an evaluation of the different versions of MEDA on a variety of challenging datasets. We compare accuracy and computational efficiency of MEDA and BOW on two datasets built for the scene recognition task. We then test the effectiveness of the two approaches for a video retrieval system built for the TRECVID 2010 [9] database.

4.1 Scene Recognition Task

We evaluate the performances of our model for image recognition in two challenging datasets, for indoor and outdoor scene categorization. First, we extract the image local descriptors using the PCA-SIFT method described in [3], which reduces the dimensionality ($d = 36$) of the original SIFT (as proposed in [4], with $d = 128$) by applying PCA on the gradient image around the salient point. Once the local descriptors are extracted, we aggregate them using both BOW, by clustering a subset of training images using a standard k-means algorithm, and MEDA models (we implement the three different versions of the MEDA model according to the methods in Sec. 3). A one vs all SVM is indeed built to separate each class from the others, using a chi-square kernel of degree 2. As evaluation measure, we use the average multiclass prediction accuracy.

Experimental Setup

The **Outdoor Scenes Dataset**, first introduced in [6], is composed of 2600 color images organized in 8 categories of natural scenes. We split such dataset using 100 images per class for training and the rest for testing. For evaluation, we compute: MEDA uniform quantization, 20 bins (*uniform* 20×36), quantile-based, 20 bins (*quantile* 20×36). entropy-based, 16 bins (*tree* 16×36). In order to compare the performances with MEDA, we create a set of visual dictionaries with 500/720 visual words (BOW 500/BOW 720). The **Indoor Scenes Dataset**,with 67 categories and 15620 images, was proposed in [7] as a new, unique database for indoor scene recognition, . For this second group of experiments, we follow the experimental setup in [7]: 20 images for testing and the rest for training. We define for this experiment: *uniform* 10×36, MEDA with uniform quantization, 10 bins ; *quantile* 10×36, quantile-based, 10 bins and *tree* 8×36, tree-based, 8 bins. Moreover, we build dictionaries of 360/500 visual words (BOW 360, BOW 500).

Results

As we can see from Table 1, the most complex version (*tree* $\cdot \times 36$) of the MEDA model is more than 150 times less computationally expensive compared to the BOW model corresponding to the same feature size. Moreover, we show in Fig 3(a-b) that MEDA is not only efficient, but it outperforms in accuracy the BOW model by 10% for the Indoor Scenes Dataset and 3% for the Outdoor Scenes.

4.2 Video Retrieval Task

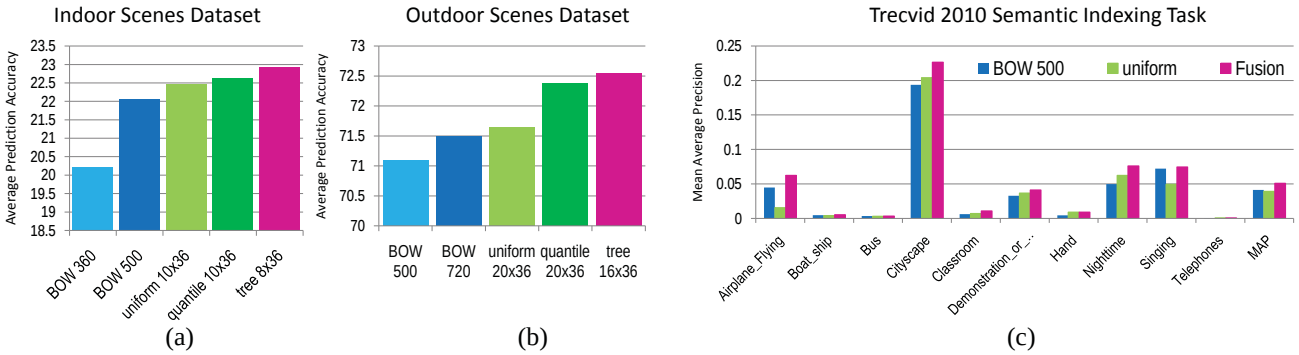

Figure 3: Comparing BOW with MEDA: (a-b) scene recognition (c) video retrieval

Table 1: Training Set Processing Times (s) for BOW vs MEDA models

Indoor 67 Dataset			Outdoor 8 Dataset		
BOW	360	62503	BOW	500	10027
	500	86685		720	14429
MEDA	uniform (10×36)	135	MEDA	uniform (20×36)	55
	quantile (10×36)	163		quantile (20×36)	23
	tree (8×36)	401		tree (16×36)	87

For the Light Semantic Indexing Task of TRECVID (SIN), participants are required to build a retrieval system that produces a ranked list of relevant shots ten semantic concepts.

Experimental setup

The training set of TRECVID 2010 contains 3200 videos that we split in 1617 for training and 1616 for test. For our experiments, we extract SIFT [4] descriptors from the video keyframes and quantize them using BOW with 500 words (*BOW 500*), as in [8], and MEDA with uniform quantization (*uniform*)[4]. A set of SVM-based classifiers is trained to detect the concept presence, for each concept. The concepts score of MEDA and BOW are then linearly fused (*fusion*) to evaluate the effectiveness of the combination of the two approaches. As evaluation measure, we use the mean average precision.

Results

Despite its simplicity and efficiency, MEDA achieves retrieval results comparable with the BOW model, as shown in Fig.3(c). The most interesting result here is the improvement we obtain on the BOW-based retrieval (+25% on the final MAP) by combining it with the MEDA-based retrieval. Our technique for descriptor aggregation brings new, complementary information to a traditional BOW model. As a matter of fact, MEDA calculates the frequency of each component (visual *letter*), while BOW calculate the frequency for the whole vector (visual *word*). Similar explanation can justify the poor results we obtain with MEDA for some concepts e.g. Airplane_Flying.

5. CONCLUSIONS

Bag of Words is an effective method for image description. Despite its accuracy, one of the major drawbacks of such model is its high computational cost. In this paper we introduced a new model for local descriptor aggregation based on descriptor marginal distribution approximation, namely the MEDA model, and demonstrated that it represents a fast and reliable approach for image categorization. Moreover, we showed that its combination with BOW improve the performances of a video retrieval system by 25 %, making MEDA a promising cue for content based multimedia retrieval. Possible tracks for future work include the usage of base LIDs other than SIFT, and the addition spatial information in the MEDA model. Moreover, a possible solution against the high dimensionality of the final image signature could be the use of techniques such as PCA. Finally, given the good performances of our 1-d quantization technique when combined with k-d quantization, we could explore the possibility of 2 or 3-d quantization and build multi-dimensional visual dictionaries.

[4]The number of bins is optimeized per concept in the training phase

6. REFERENCES

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22. Citeseer, 2004.

[2] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *In Proceedings of the IEEE International Conference on Computer Vision.*

[3] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE conference on Computer vision and pattern recognition.*

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[5] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2. IEEE, 2006.

[6] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[7] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, IEEE Conference on.* IEEE, 2009.

[8] M. Redi, B. Merialdo, and F. Wang. Eurecom and ecnu at trecvid 2010: The semantic indexing task. In *In Proceedings of the TRECVID 2010 Workshop*, 2010.

[9] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06*, New York, NY, USA, 2006. ACM Press.

[10] T. Tuytelaars. Vector quantizing feature space with a regular lattice. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.*