Exploring Two Spaces With One Feature: Kernelized Multidimensional Modeling of Visual Alphabets

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

Marginal Alphabets (MEDA) were proposed as an alternative to Bag of Words (BoW) for image representation. They aggregate sets of locally extracted descriptors (LEDs) by using visual alphabets based on the marginal approximation of the LED components. Compared to the *exponential* complexity of the BoW codebooks, the MEDA model is very efficient because each dimension of the LED is quantized independently. However, MEDA lacks of considering the relations between the LED components, loosing precious information for image representation.
In this paper, we design Multi-MEDA, a shift-invariant kernel for MEDA signatures that allows to reintroduce, at a kernel level, the connections between LED components that were broken with the independent quantization. With our approach, we can derive in a *polynomial* time a multivariate model from the marginal approximations stored in the MEDA vector, without explicitly computing any multidimensional codebook. Results show that the MEDA signature increases its discriminative power when analyzed through the Multi-MEDA kernel evaluation. Moreover, we show that the model generated my the Multi-MEDA-based learning brings complementary information compared to traditional kernels over MEDA and BoW signatures: our experiments on the TRECVID database show that the combination of these approaches brings a substantial improvment compared to BoW-only classification.

Categories and Subject Descriptors

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

Keywords

Scene Recognition, Feature Extraction, CBIR

1. INTRODUCTION

Visual signatures represent a crucial element for the development of effective Content-Based Multimedia Retrieval (CBMR) and categorization systems. In CBMR frameworks, signatures are extracted from visual data, and then modeled using kernel-based learning techniques such as Support Vector Machines. Image signatures based on aggregation of locally extracted descriptors (LEDs) such as SIFT [9] have been proved to be effective representations for CBMR when associated with kernel machines.

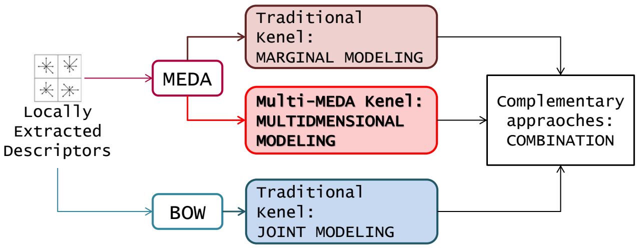


Figure 1: Multi-MEDA: a kernel over MEDA descriptors for multivariate probability modeling from marginals

One of the most popular LED-based model is probably the Bag of Visual Words model (BoW). In this approach, LEDs, namely local feature vectors of fixed length k, are first computed to describe the surrounding of interest points [4] or densely sampled points [3]. A codebook of w visual words is then generated by vector quantizing the k-dimensional space defined by the LED. Each image is then mapped into a w-dimensional signature by collecting the occurrences of such words. In order to avoid information loss during this mapping, the codebook needs to properly reflect the joint distribution of the LED components: the codebook creation is therefore generally performed by quantizing the LEDs with clustering techniques such as k-means [2] or hierarchical clustering [10]. Despite its effectiveness, a negative impact of the BoW model is its associated complexity and storage cost, that increase *exponentially* with the length of the LED.

An alternative and complementary approach for LED aggregation is represented by the MEDA [12] signature. In this approach, the visual codebook is a set of n uni-dimensional bins ("letters") per dimension, obtained by quantizing the marginal distribution of each component of the LED. The final image representation is a $k \times n$ histogram collecting the occurrences of such letters at each dimension. The MEDA signature represents therefore a concatenation of the approximated marginal distributions of the image LEDs components. This approach is very efficient, because it performs the vector quantization in a 1-d space, eliminating the correlation between the LED components by analyzing their distributions independently. However, by doing so, MEDA

ICMR '12, June 5-8, Hong Kong, China

brakes the relationships between the LED components, loosing a lot of useful information for image representation.

In this paper we propose Multi-MEDA, namely a new kernel function designed for the MEDA signature, that allows to model the joint contribution of the LED components in an efficient way. Our Multi-MEDA kernel models in a *polynomial* time the k-dimensional LED space, by deriving a multivariate probability from the k marginal approximations. With our approach, we keep as input to the kernel machine the classic, marginal-based MEDA signature, but we increase its discriminative power by analyzing it under a multidimensional perspective through the kernel formulation.

The main idea behind the Multi-MEDA kernel is that, since MEDA considers each dimension of the LED as an independent variable, we can approximate the joint distribution of the LED components by multiplying their marginal distributions. However, an image signature supporting such model would require a k-fold cartesian product of n-dimensional vectors, namely the multiplication of the approximations concatenated in the MEDA signature. This would lead again to an exponentially complex problem ($O(n^k)$), with a codebook of n^k elements and an extremely high-dimensional feature. For this reason, the key aspect of our approach is that we do not compute explicitly the image signature nor the visual dictionary, and instead we shift the computation of the multivariate probability inside the kernel machine. As a matter of fact, Multi-MEDA is as shift-invariant kernel that embeds the marginals multiplication, i.e. the cartesian product of the marginal approximations. The most important property of our kernel is that it does not require exponential time to achieve the multidimensional modeling. We indeed show that the cost of computing the k-dimensional joint probability with the Multi-MEDA kernel becomes polynomial with the dimension of the LED and the number of letters in the MEDAcodebook (($O(nk)$). Therefore, although MEDA is built to describe marginal 1-d probabilities, when place the Multi-MEDA kernel on top of MEDA signatures, we can reconsruct a model of the LED space that is based on a $k - d$ multi-variate probability, without needing to quantize the $k - d$ space.

Compared to the models generated by traditional SVM kernels over MEDA signatures, Multi-MEDA represents the LED space under a new, complementary point of view, as shown in Fig. 1. Multi-MEDA allows therefore to explore two spaces (marginal-based and multidimensional) with the same feature (MEDA). Moreover, both the MEDA model and Multi-MEDA model are in turn different from the joint distribution approximation generated by traditional BoW approaches. By introducing Multi-MEDA, we therefore introduce a new discriminative source of information regarding the LED distribution, that can be combined with the MEDA and BoW models, leading to a significant increase (+50 %) of the CBMR performances, without requiring the computation of new LEDs, and without introducing exponential complexity. We test the effectiveness of our solution for scene recognition and video retrieval on a variety of challenging datasets (e.g. the TrecVid [14] 2010 database), and show that the Multi-MEDA model achieve good performances for all the tasks. Moreover, we show that The combination of MEDA, Multi-MEDA and BoW produces substantial improvements (more than 50% on the TrecVid data) compared to BoW-only retrieval.

The rest of the paper is organized as follows: in Sec. 2 we give an overview of the related work, Sec. 3 gives a statistical explanation of the differences and complementarities between MEDA, Multi-MEDA and BoW. In Sec. 4 we give a brief overview of the MEDA model, and show its view from a kernel perspective. In Sec. 5 we detail the implementation of our new kernelized solution and finally in Sec. 6 we validate the proposed framework with experimental results on scene categorization and video retrieval.

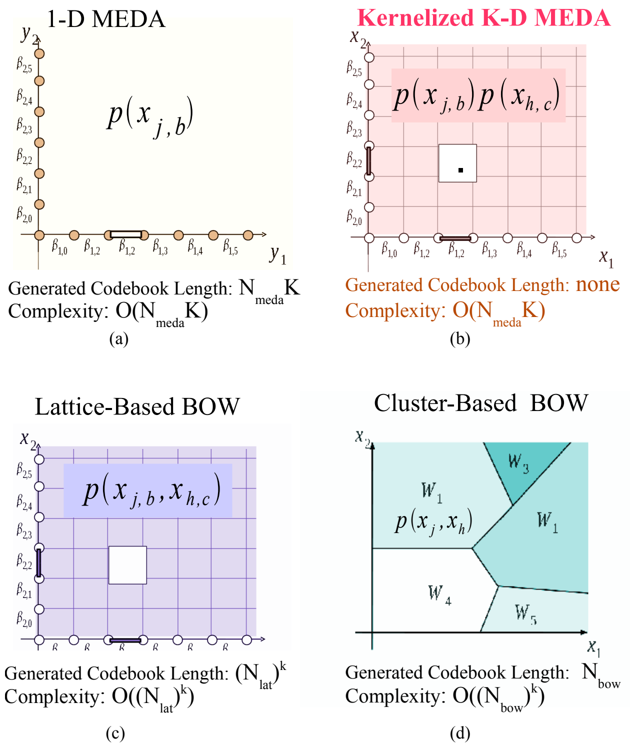
2. RELATED WORK

In Multi-MEDA we look at the cooperation between LED-based signatures and kernels, and we shift some statistical aspects of the features at a kernel level. In this section, we will summarize the relevant work in the field that directly relates to our proposed approach. First, we will show the most significant contributions in the area of the LED-based image representation. We will then give an overview of the state-of-the art works that focus on the combination between LED aggregators and kernels to improve the discriminative ability of the image signature.

The popular BoW model is the most used framework for image representation based on locally extracted descriptors. A major issue in the BoW model is the way in which the LED quantization in the k-d space is performed, namely, the method used to define the visual dictionary. Csurka et al. in [2] first introduced the BoW approach, addressing the quantization problem by applying k-means clustering on the LEDs of the training set. This approach was then extended in [17] to select the optimal set of words based on a discrimination measure. Various clustering techniques have been used later on to vector quantize the LED space. For example, to create visual codebooks, in [8] a mean-shift clustering is proposed to obtain visual words from LEDs and in [10] LEDs are hierarchically quantized in a vocabulary tree. An alternative approach is represented by [15], where each dimension of the LED is quantized into a fixed set of bins, and then the dictionary is chosen based on the discriminative power of the resulting hypercubes. All the mentioned approaches are proved to be very effective for image categorization and retrieval tasks. However, one of the major drawbacks is that they perform the codebook computation in the k-dimensional space, which implies exponential computational time with the number of dimensions of the LED.

Another view is given by the MEDA model: a 1-dimensional search approach was proposed in [12], that quantizes each dimension of the LED into a fixed number of bins, and obtains the image signature by counting the occurrences of such bins in the components of the image LED. This approach is proved to be very light in terms of computation, and as efficient as BoW for moderate dictionary sizes. Since the quantization is performed independently for each dimension, MEDA does not take into account the relationships between the behavior of the single LED components. However, LED are vectors that describe an entire image region, and each element in a LED concur in describing the surroundings of an interest point. Therefore, eliminating the correlation between the LED elements can cause losses of precious information for image description.

This motivates us to perform a kernel-based analysis on the MEDA signature, allowing to learn a multivariate model that considers the relations between the LEDs components. To our knowledge, the Multi-MEDA approach is one of the


1-D MEDA
y_2
β_{25} β_{24} β_{23} β_{22} β_{21} β_{20}
$p(x_{j,b})$
$\beta_{1,0}$ $\beta_{1,2}$ $\beta_{1,2}$ $\beta_{1,3}$ $\beta_{1,4}$ $\beta_{1,5}$
y_1
Generated Codebook Length: $N_{meda}K$
Complexity: $O(N_{meda}K)$
Kernelized K-D MEDA
x_2
$p(x_{j,b})p(x_{h,c})$
x_1
Generated Codebook Length: none
Complexity: $O(N_{meda}K)$
Lattice-Based BOW
x_2
$p(x_{j,b}, x_{h,c})$
Generated Codebook Length: $(N_{lat})^k$
Complexity: $O((N_{lat})^k)$
Cluster-Based BOW
x_2
W_1
$p(x_j, x_h)$
W_3
W_1
W_4
W_5
x_1
Generated Codebook Length: N_{bow}
Complexity: $O((N_{bow})^k)$

(a)

(b)

(c)

(d)

Figure 2: Statistical and time complexity differences between various LDE aggregators: (a) MEDA, (b) Multi-MEDA (c) Lattice-based BOW (d) Clustering-based BOW.

first attempts to improve the MEDA model by focusing on the kernel properties.

As a matter of fact, the cooperation between LED aggregators and kernels has mainly been investigated for extending the traditional BoW model. Various steps of the BoW approach has been improved through the interaction with kernels: a better codebook generation is achieved in [18] by using the Histogram Intersection Kernel in an unsupervised manner, while in [5] codebooks are used as free parameters of a Multiple Kernel Learning-based learning algorithm. In [16], kernel density estimation techniques are used to model ambiguities in visual word assignment. Lazebnik et al. in [7] use Spatial Pyramid Kernels to add the spatial information in the BoW model learning. The learning step is also improved in [1] by mapping the image LEDs into a low dimensional feature space, then averaging such vectors to obtain a set-level feature, and finally using a linear classifier to model the resulting vectors.

Our approach is different from the mentioned approaches because, first of all, we do not analyze the BoW model, but we extend instead the MEDA model to a multi-dimensional model through a kernel-based learning. Moreover, although we generate a model that works on a multivariate probability, the space that we explore through the Multi-MEDA kernel; is statistically different from the space determined by vector quantization in BoW. We will see in the next section the reasons of this difference.

3. FEEDING 2 BIRDS WITH ONE SEED: SAME FEATURES, 2 (3) SPACES

The kernelized approach proposed in this paper represents a new technique to approximate the multivariate probability distribution of the LEDs in an image dataset. As mentioned, various approaches [2, 12, 15] are available in literature to produce compact image signatures based on the distribution of the LEDs. In this section, we will show the novelty introduced by Multi-MEDA with respect to existing techniques and demonstrate that our method is complementary to the state-of-the-art LED aggregators from a statistical point of view .

3.1 Multi-MEDA VS Existing LED-based signatures

Here we analyze statistically the BoW [2] and the MEDA [12] models and compare them with the Multi-MEDA approach. For simplicity, we assume a 2-d space, e.g. an ideal case where the length of the LED is $k = 2$. A visual representation of our analysis can be seen in Figure 2. We define as $X = (x_1^i, x_2^i)$, $i = 1, \dots, t$ the set of t descriptors extracted from a given image set.

In the BoW model(see Fig. 2(d)), vector quantization is performed in the LED k-dimensional space using a variety of approaches [2, 8, 15]. No matter the approach used, the output is a codebook of N_{bow} vectors $c^{bow} = \{c_1^{bow}, \dots, c_{N_{bow}}^{bow}\}$ that allows to approximate the joint distribution of the LED components, namely the multivariate $p^{bow} = p(x_1, x_2)$. Lattice-based models like [15] (see Fig. 2(c)) follow the same model: although each dimension is quantized into a fixed number of bins N_{lat}, The vocabulary is obtained by generating all the resulting N_{lat}^k hypercubes, and then discarding the less discriminative ones based on the conditional joint distribution of the LED components.

In the MEDA model, as shown in Fig. 2 (a), each dimension of the LED is quantized independently, therefore the codebook is a set of N_{md} letters for each dimension,
$c^{md} = (c_{1,1}^{md}, \dots, c_{1,N_{md}}^{md}, \dots, c_{2,1}^{md}, \dots, c_{2,N_{md}}^{md})$,
that reflects the marginal of each dimension, namely the 1-dimensional $p_1^{md} = p(x_1)$, or $p_2^{md} = p(x_2)$.

In our Multi-MEDA model, Fig. 2(b), we exploit the independence between the LEDs dimensions introduced by MEDA, and we estimate the joint LEDs distribution by multiplying the 1-d marginals, therefore $p^{mmd} = p_1^{md} \cdot p_2^{md} = p(x_1) \cdot p(x_2)$. Since the computation of this multivariate distribution is performed inside the Multi-MEDA kernel, in our approach we do not need to compute a new visual signature or express explicitly the shared codebook, and we use instead as input the traditional MEDA vector.

3.2 1 Signature, 2 Spaces. 1 LED, 3 spaces.

p^{md} and p^{mmd} are therefore generated using the same feature vector, but analyzing it with different kernels (traditional RBF or linear in the first case, Multi-MEDA in the second case). However, while the first one is a 1-d, marginal probability, the second is an actual multivariate probability distribution. Therefore, with our approach, we allow to construct two different models of the LED space using the same input vector. The two models generated represent different sources of information regarding the position of the examples in the feature space. In this way, we “feed two birds with one seed”: we explore two, complementary, probability distributions using one single descriptor.

Furthermore, if we consider the LED as the input "seed", we are in a way actually feeding "three birds". As a matter of fact, p^{bow} and p^{mmd} are both multi-dimensional approximations of the LED distribution. However, while the first one represents an estimation of the real joint probability, the second is a k-d probability inferred from the set of k, monodimensional p^{md}. We can therefore say that also p^{bow} and p^{mmd} allow to learn the LED space with different, complementary approaches. Moreover, it was already proved in [12] that MEDA and BoW represent orthogonal approaches to aggregate LEDs. We can therefore deduce that the three approaches discussed (BoW, MEDA and Multi-MEDA) represent three different cues to explore the LED space, and that we can therefore combine their contributions in order to maximize the CBMR performances.

4. MATCHING MEDA WITH KERNELS: MARGINAL APPROXIMATIONS

In order to understand the Multi-MEDA approach, we detail in this section the implementation of the MEDA signature and show its kernel perspective, namely how the kernel function is formulated when evaluating MEDA vectors.

4.1 Marginals Estimation for Descriptors Aggregation

In the MEDA approach, each dimension of the LED is quantized in a set of n bins, namely the "letters" of the visual alphabet. The final image signature is then obtained by collecting the number of feature vectors that fall into a given bin, for each dimension.

Like traditional BoW models, for an image I, m salient points are detected in the image. For each point, a k-dimensional normalized SIFT descriptor $x^i = (x^i_1, \dots, x^i_k)$, $i = 1, \dots, m$ is then computed to describe its surrounding region. Each component x_j, $j = 1, \dots, k$ of the descriptor is then quantized into n discrete values $\beta_{j,b}$, $b = 1, \dots, n$ according to its marginal distribution $p(x_j)$. A set of k independent alphabets $c_1, \dots, c_k$ results from this quantization, where $c_j = \beta_{j,1}, \dots, \beta_{j,n}$.

The final image representation is a $k \times n$ histogram

$$v = (v_{1,1}, v_{2,1}, \dots, v_{n,1}, v_{1,2}, \dots, v_{k,n}). \quad (1)$$

Each element $v_{j,b} = p(x_{j,b}) = \#\{x^i : x^i_j \in \beta_{j,b}\}$ in the MEDA signature counts how many image descriptors at position j fall into bin b.[1]

4.2 MEDA from a kernel perspective

In retrieval frameworks, kernel machines learn the input space using as input visual descriptors such as MEDA. In the learning phase, the machine learns how to separate the feature space into two classes. In order to do so, kernels are used to evaluate similarities between such features and define an optimal decision boundary, namely a hyperplane in the feature space.[2]

Among the many kernel function used to model the feature space (e.g. chi-square, polynomial), the Radial Basis Function (RBF) kernel has been shown to perform well for image retrieval applications [19].

For two input vectors v and w, the RBF kernel has equation

$$k(v, w) = \exp(-\lambda||v - w||^2).$$

When MEDA is used in conjunction with a RBF-based classifier, the kernel function evaluates the differences between the letters frequencies for each pair of training images I and J. In order to show this behavior, We will use the following notation:

- for image I, the LEDs are in the set $x = \{x^i_j\}$ and the MEDA signature is $v = \{v_{j,b}\}$
- for image J, the LEDs are in the set $y = \{y^i_j\}$ and the MEDA signature is $w = \{w_{j,b}\}$

In order to understand the kernel view of MEDA, in Figure 3, we propose a 2-d representation (namely a scenario where the LED has dimension $k = 2$) of the MEDA-based feature space. In the 2-dimensional case, the kernel function of MEDA signatures becomes:

$$\begin{aligned} k(v, w) &= \exp(-\lambda(\sum_{b=1}^{n}|v_{1,b} - w_{1,b}|^2 + \sum_{b=1}^{n}|v_{2,b} - w_{2,b}|^2) \\ &= \exp(-\lambda(\sum_{b=1}^{n}|p(x_{1,b}) - p(y_{1,b})|^2 + \\ &\quad + \sum_{b=1}^{n}|p(x_{2,b}) - p(y_{2,b})|^2)), \end{aligned} \quad (2)$$

i.e. for each dimension j, the sum over n bins of the squared differences between the signature values at each bin b.

It is therefore straight-forward to extend such kernel view and consider the real case, i.e. when $k >> 2$. In this scenario, the kernel evaluates the marginal contribution of all dimensions ($j = 1, \dots, k$) and the previous equation becomes:

$$k(v, w) = \exp(-\lambda(\sum_{j=1}^{k}\sum_{b=1}^{n}|p(x_{j,b}) - p(y_{j,b})|^2)) \quad (3)$$

As confirmed by the summation of Eq. (3) the current formulation of the MEDA signature analyzes the marginal distribution of each dimension of the LED independently, without taking into account the interactions between the components in the k-dimensional space.

5. KERNELIZED MULTI-MEDA: MULTIDIMENSIONAL PROBABILITY ESTIMATION FROM MARGINALS

As explained before, the MEDA modeling generates a 1-dimensional probability, while a model based on LED k-dimensional vectors should exploit a multivariate probability. In the Multi-MEDA model, we derive a k-dimensional probability from the marginal (1-d) probabilities computed for each dimension of the LED. Since the computation of such signature would result in an extremely high-dimensional vector, we shift the multidimensional modeling at a kernel

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

[2] When the input samples are linearly separable, the similarity between two features v and w is computed with a simple dot product $v \cdot w$. However, in many cases, e.g. in multimedia data representation, decision boundary is not linear: one common solution is to define a transform ϕ that maps the input space in the feature space $v \to \phi(v)$ and then use a kernel function $k(v, w) = \phi(v) \cdot \phi(w)$ to represent the dot product in the high-dimensional feature space.

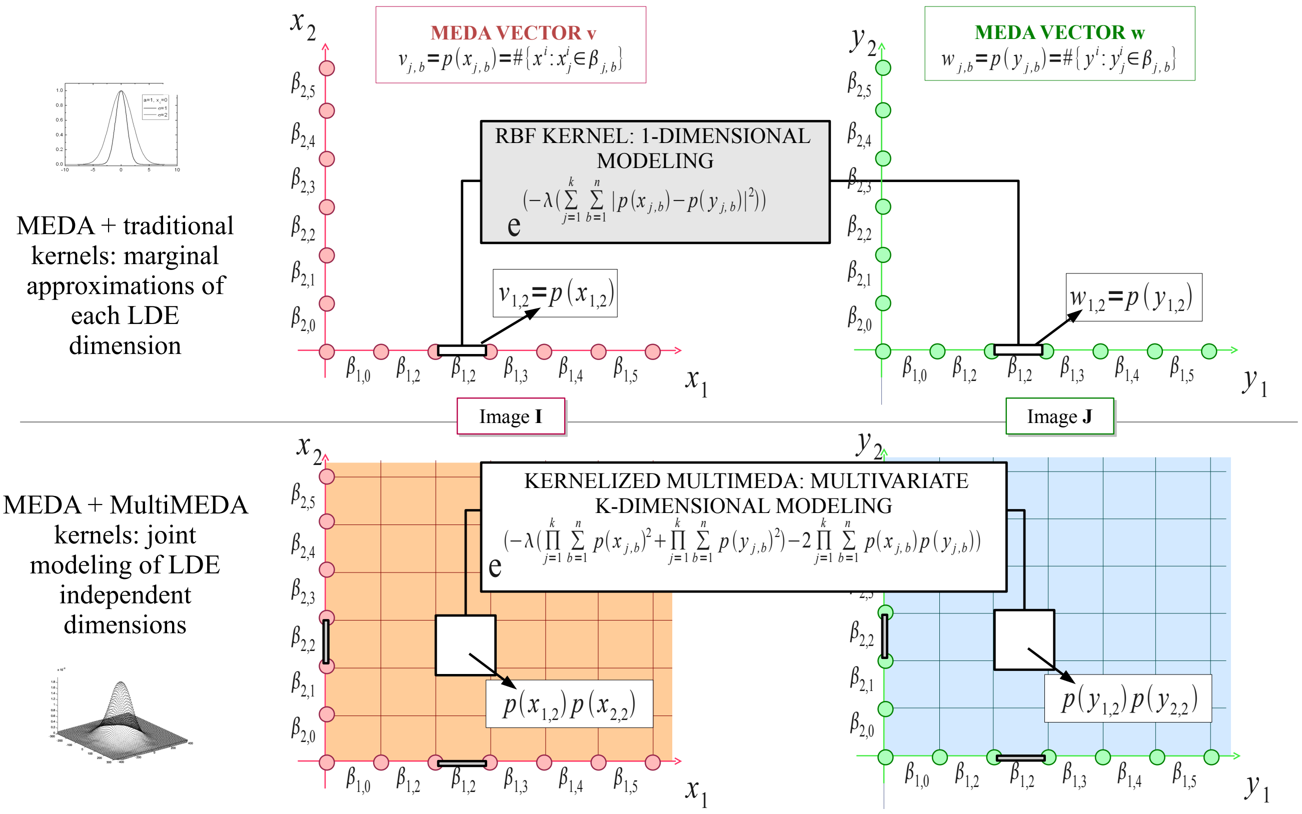

MEDA + traditional kernels: marginal approximations of each LDE dimension
MEDA VECTOR v
$v_{j,b}=p(x_{j,b})=\#\{x^i : x^i_j \in \beta_{j,b}\}$
MEDA VECTOR w
$w_{j,b}=p(y_{j,b})=\#\{y^i : y^i_j \in \beta_{j,b}\}$
RBF KERNEL: 1-DIMENSIONAL MODELING
$e^{(-\lambda(\sum_{j=1}^{k}\sum_{b=1}^{n}|p(x_{j,b})-p(y_{j,b})|^2))}$
$v_{1,2}=p(x_{1,2})$
$w_{1,2}=p(y_{1,2})$
x_1
x_2
y_1
y_2
$\beta_{1,0}$ $\beta_{1,2}$ $\beta_{1,2}$ $\beta_{1,3}$ $\beta_{1,4}$ $\beta_{1,5}$
$\beta_{2,0}$ $\beta_{2,1}$ $\beta_{2,2}$ $\beta_{2,3}$ $\beta_{2,4}$ $\beta_{2,5}$
Image **I**
Image **J**
MEDA + MultiMEDA kernels: joint modeling of LDE independent dimensions
KERNELIZED MULTIMEDA: MULTIVARIATE K-DIMENSIONAL MODELING
$e^{(-\lambda(\prod_{j=1}^{k}\sum_{b=1}^{n}p(x_{j,b})^2+\prod_{j=1}^{k}\sum_{b=1}^{n}p(y_{j,b})^2)-2\prod_{j=1}^{k}\sum_{b=1}^{n}p(x_{j,b})p(y_{j,b}))}$
$p(x_{1,2})p(x_{2,2})$
$p(y_{1,2})p(y_{2,2})$

Figure 3: Placing kernels on top of MEDA: marginal (RBF/traditional kernels) and multidimensional (Multi-MEDA kernel) approaches

level, embedding the k-d evaluation of MEDA in a RBF kernel. In this section, we start from the MEDA formulation and show its multi-dimensional extension, that we then kernelize to build the Multi-MEDA model.

The MEDA vector in Eq. 1 can be seen as a set of k n-dimensional vectors $\{v_j\} = \{p(x_{j,1}), \dots, p(x_{j,n})\}$, $j = 1, \dots, n$. Each v_j represents the approximation of the marginal $p(x_j)$ of the j^{th} component of the LEDs in image I. We want to derive a joint probability so that we can exploit the combination of the occurrences of all the dimensions. Since MEDA analyzes each dimension independently, in order to estimate the joint probability, we can multiply the contribution of the marginals of all components.[3]

For image I, this would result in a k-dimensional vector determined by the k-fold cartesian product of all vectors v_j, $\forall j$. The model codebook would be the cartesian product of all the k scalar alphabets, namely the set of hypercubes:

$$\begin{aligned} c_{(1,2,\dots,k)} &= c_1 \times c_2 \times \dots, \times c_k = \\ &= \{(\beta_{1,1}, \dots, \beta_{k,1}), \dots, (\beta_{1,n}, \dots, \beta_{k,n})\} = \\ . &= \{(\beta_{1,b}, \beta_{2,d}, \dots, \beta_{k,e}\},\ b, d, e = 1, \dots, n \end{aligned}$$

Each value of the n^k-dimensional signature would be therefore the product of the occurrences of the unidimensional bins that concur in generating each hypercube:

$$v_{(1,b),(2,d),\dots,(k,e)} = p(x_{1,b}) \cdot p(x_{2,d}) \cdot \dots \cdot p(x_{k,e}). \qquad (4)$$

The number of hypercubes to consider in such multidimensional formulation of the MEDA signature is exponential with the number of dimensions of the LED, which is typically 128 for traditional SIFT [9] vectors or 36 for PCA-SIFT [6]. Treating such high-dimensional feature, even with a small number of training samples, becomes impractical with traditional kernel machines. This motivates us to shift this multivariate probability computation inside an RBF-like kernel, and create the Multi-MEDA kernel.

As proposed for the previous analysis, we start with the 2-d example ($k = 2$, see Figure 3) and we then extend it to the more realistic k-d case.

When we want to take the cartesian product of the marginals (as in Eq. (4) when $k = 2$) inside an RBF-like kernel, for the two images I and J the formulation in Eq. 3 becomes

$$k(v, w) = \exp(-\lambda(\sum_{b=1,c=1}^{n} |p(x_{1,b}) \cdot p(x_{2,c}) - p(y_{1,b}) \cdot p(y_{2,c})|^2)) \qquad (5)$$

Developing the power in Eq. (5), we obtain:

$$\begin{aligned} k(v, w) &= \exp(-\lambda(\sum_{b=1,c=1}^{n} p(x_{1,b})^2 \cdot p(x_{2,c})^2 + p(y_{1,b})^2 \cdot p(y_{2,c})^2 - \\ &\quad -2(p(x_{1,b}) \cdot (y_{1,b}) \cdot (x_{2,b}) \cdot (y_{2,c})))) \\ &= \exp(-\lambda(\sum_b p(x_{1,b})^2 \sum_c p(x_{2,c})^2 + \sum_b p(y_{1,b})^2 \sum_c p(y_{1,c})^2 \\ &\quad -2 \sum_b p(x_{1,b}) p(y_{1,b}) \sum_c p(x_{2,c}) p(y_{2,c}))). \qquad (6) \end{aligned}$$

The trick that allows us to compute Multi-MEDA in a polynomial time is that, when extending Eq. (6) to the k-dimensional space, the squares of the MEDA elements are

[3] $(P(A, B) = P(A) \cdot P(B)$ if A, B are independent)

multiplied over all dimensions independently, and the previous Equation becomes:

$$k(v, w) = \exp(-\lambda(\prod_{j=1}^{k}\sum_{b=1}^{n} p(x_{j,b})^2 + \prod_{j=1}^{k}\sum_{b=1}^{n} p(y_{j,b})^2 - \\ -2\prod_{j=1}^{k}\sum_{b=1}^{n} p(x_{j,b})p(y_{j,b}))), \quad (7)$$

which has a polynomial complexity $O(kn)$. This allows us to use directly the original MEDA vectors as input to the kernel-based classifier, without pre-computing the dictionary hypercubes and the multidimensional MEDA (Eq.(4)). Moreover, unlike [15], this product-based formulation allows us to increase both the number of letters in the 1-d alphabets and the LED dimension without exponential increase of computation.

6. EXPERIMENTAL VALIDATION

We tested the effectiveness of the kernelized Multi-MEDA on two recognition tasks, namely scene categorization and video retrieval. We will use the following naming convention here: MEDA is the MEDA signature learnt with traditional kernels, Multi-MEDA is the MEDA signature learnt with the Multi-MEDA kernel, and BoW represents the Bag of Words vector. We compute the three models on the input images and we compare their performances. We use them as stand-alone descriptor and we then analyze the effects of their combinations, on the two given tasks. In this section, we show that our proposed multidimensional modeling achieves good performances in both the mentioned tasks, comparable with both MEDA and BoW. Moreover, when we combine Multi-MEDA with the other LED aggregators, we show that it actually provides complementary information, as hypotized in Sec.3, bringing a significative improvement in our experimental results.

MEDA, BoW, and Multi-MEDA share the same input seed. Therefore, the first step of our experiments is the extraction of a set of SIFT keypoints. We then aggregate them using both BoW, by clustering a subset of training images using a standard k-means algorithm, and MEDA models (using the percentile-based technique proposed in [12]). Both signatures are then learnt by chi-square kernels.

In order to compute the kernelized Multi-MEDA, we use the MEDA signatures as input for our RBF-based multidimensional kernel in Eq. (7). One major issue is that the MEDA values are not normalized. Therefore, the products over k dimensions in Eq. (7) result in very high values. This values become the negative exponent of the RBF kernel, and $k(v, w)$ becomes close to zero. The similarity between the two vectors cannot be estimated reliably without normalization. In order to cope with this issue, we normalize the MEDA signature by m/n inside the kernel formulation. This is because each element in the MEDA vector represents a fraction (approximately $1/n$) of the total number of vectors (m), namely the one that take a given value in a given dimension. Moreover, instead of taking the product of such small values, that would bring the exponent to zero, we compute the sum of the log of those terms. [4]

In the following experiments, other parameters or vector quantization models can be used, but given the statistical difference between the three approaches, the performances. of the stand-alone models and their combined contributions would not change significantly.

6.1 Scene Categorization

For the task of scene categorization, we choose two different datasets, namely the Indoor-67 database [11], and the Amadeus-16 database used in [13]. For each database, we select a different experimental setup, and we look at the experimental results:

Amadeus-16 dataset: travel-related scene categorization

The Amadeus dataset is a set of 100,000 hotel-related images coming from a travel service provider, that are used in Hotel Management Platforms. This database contains 16 indoor and outdoor scene categories, annotated from different sources and therefore particularly subject to labeling noise. For each category, half of the images are used for training and the rest for testing. For this group of experiments, we extract PCA-Sift LEDs [6]. We then compute the following signatures

- BoW with 150 visual words
- MEDA with 5 letters per dimension (total signature length is 180)

Both signatures are learnt with a 1-vs-all SVM with chi-square kernel. The MEDA vectors are then used as input for 1-vs-all SVM with Multi-MEDA kernel to compare the performances. The three features prediction are then combined with weighted linear fusion.

Results on this database show that actually the Multi-MEDA kernel models the LED space in a meaningful and effective way: Multi-MEDA achieves comparable results with both BoW and MEDA, and the combination of the MEDA and the Multi-MEDA contributions (same feature, different kernels) outperforms the BoW model. Moreover, we can see here someF evidences of the complementarity of the three approaches: the combination of MEDA, Multi-MEDA and BoW gives an improvement of about 6% over the MEDA-only based classification.

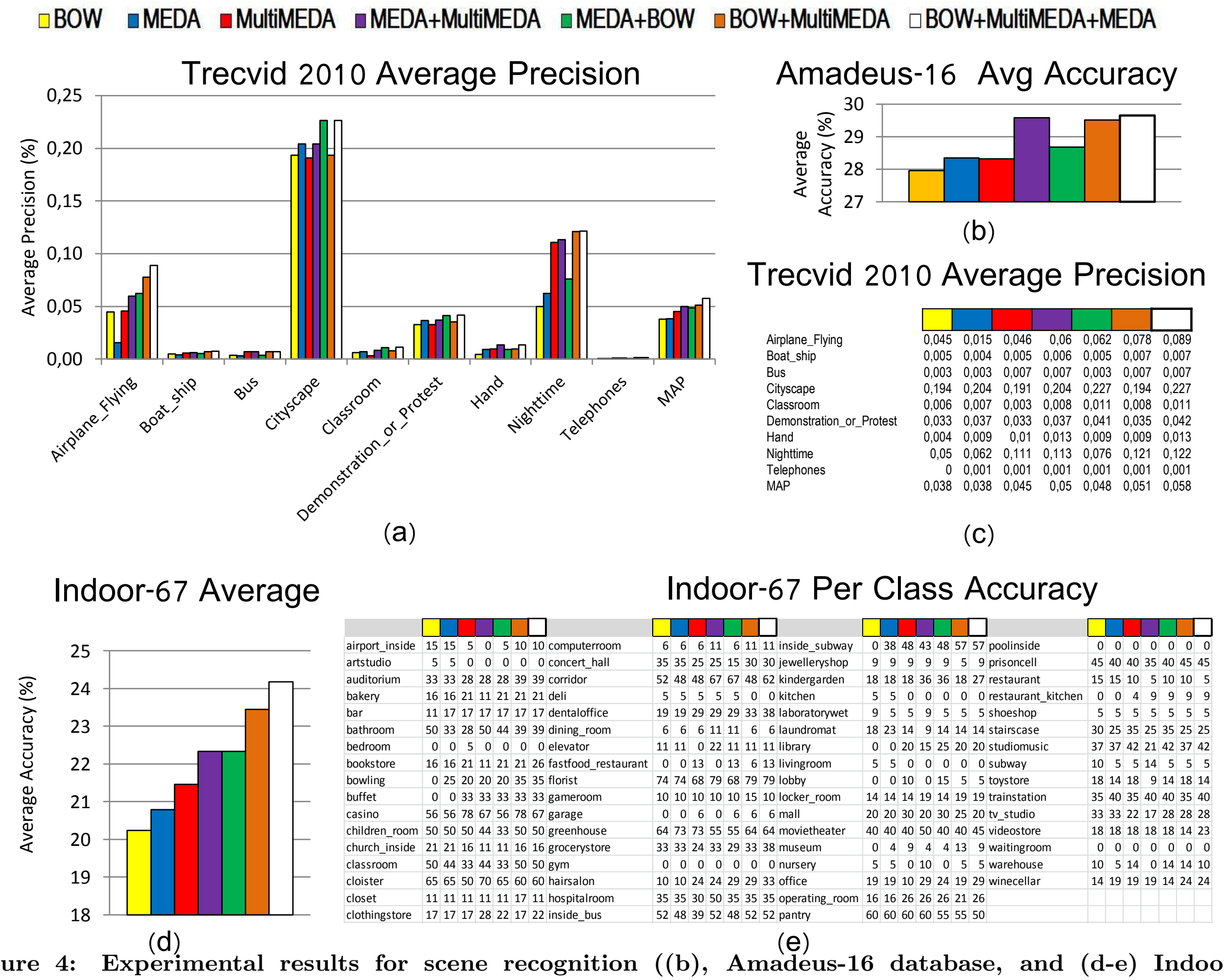
Indoor-67 dataset: indoor scene categorization

This database was first introduced in [11] for indoor scene recognition with global and local features, and it has now become a benchmarking dataset for scene categorization descriptors. It spans 67 categories with around 15500 images of different sizes. For this database, we extract PCA-Sift LEDs [6]. We then compute the LED aggregators, namely, similar to [12]:

- BoW with 360 visual words
- MEDA with 10 letters per dimension (total signature length is 360)

[4] Equation (7) becomes therefore:

$$k(v, w) = \exp(-\lambda(\sum_{j=1}^{k} log((\frac{n}{m})^2 \sum_{b=1}^{n} p(x_{j,b})^2) + \\ + \sum_{j=1}^{k} log((\frac{n}{m})^2 \sum_{b=1}^{n} p(y_{j,b})^2) \\ -2\sum_{j=1}^{k} log((\frac{n}{m})^2 \sum_{b=1}^{n} p(x_{j,b})p(y_{j,b})))). \quad (8)$$



	🟨	🟦	🟥	🟪	🟩	🟧	⬜
Airplane_Flying	0,045	0,015	0,046	0,06	0,062	0,078	0,089
Boat_ship	0,005	0,004	0,005	0,006	0,005	0,007	0,007
Bus	0,003	0,003	0,007	0,007	0,003	0,007	0,007
Cityscape	0,194	0,204	0,191	0,204	0,227	0,194	0,227
Classroom	0,006	0,007	0,003	0,008	0,011	0,008	0,011
Demonstration_or_Protest	0,033	0,037	0,033	0,037	0,041	0,035	0,042
Hand	0,004	0,009	0,01	0,013	0,009	0,009	0,013
Nighttime	0,05	0,062	0,111	0,113	0,076	0,121	0,122
Telephones	0	0,001	0,001	0,001	0,001	0,001	0,001
MAP	0,038	0,038	0,045	0,05	0,048	0,051	0,058

	🟨	🟦	🟥	🟪	🟩	🟧	⬜
airport_inside	15	15	5	0	5	10	10
artstudio	5	5	0	0	0	0	0
auditorium	33	33	28	28	28	39	39
bakery	16	16	21	11	21	21	21
bar	11	17	17	17	17	17	17
bathroom	50	33	28	50	44	39	39
bedroom	0	0	5	0	0	0	0
bookstore	16	16	21	11	21	21	26
bowling	0	25	20	20	20	35	35
buffet	0	0	33	33	33	33	33
casino	56	56	78	67	56	78	67
children_room	50	50	50	44	33	50	50
church_inside	21	21	16	11	11	16	16
classroom	50	44	33	44	33	50	50
cloister	65	65	50	70	65	60	60
closet	11	11	11	11	11	17	11
clothingstore	17	17	17	28	22	17	22

	🟨	🟦	🟥	🟪	🟩	🟧	⬜
computerroom	6	6	6	11	6	11	11
concert_hall	35	35	25	25	15	30	30
corridor	52	48	48	67	67	48	62
deli	5	5	5	5	5	0	0
dentaloffice	19	19	29	29	29	33	38
dining_room	6	6	6	11	11	6	6
elevator	11	11	0	22	11	11	11
fastfood_restaurant	0	0	13	0	13	6	13
florist	74	74	68	79	68	79	79
gameroom	10	10	10	10	10	15	10
garage	0	0	6	0	6	6	6
greenhouse	64	73	73	55	55	64	64
grocerystore	33	33	24	33	29	33	38
gym	0	0	0	0	0	0	0
hairsalon	10	10	24	24	29	29	33
hospitalroom	35	35	30	50	35	35	35
inside_bus	52	48	39	52	48	52	52

	🟨	🟦	🟥	🟪	🟩	🟧	⬜
inside_subway	0	38	48	43	48	57	57
jewelleryshop	9	9	9	9	9	5	9
kindergarden	18	18	18	36	36	18	27
kitchen	5	5	0	0	0	0	0
laboratorywet	9	5	5	9	5	5	5
laundromat	18	23	14	9	14	14	14
library	0	0	20	15	25	20	20
livingroom	5	5	0	0	0	0	0
lobby	0	0	10	0	15	5	5
locker_room	14	14	14	19	14	19	19
mall	20	20	30	20	30	25	20
movietheater	40	40	40	50	40	40	45
museum	0	4	9	4	4	13	9
nursery	5	5	0	10	0	5	5
office	19	19	10	29	24	19	29
operating_room	16	16	26	26	26	21	26
pantry	60	60	60	60	55	55	50

	🟨	🟦	🟥	🟪	🟩	🟧	⬜
poolinside	0	0	0	0	0	0	0
prisoncell	45	40	40	35	40	45	45
restaurant	15	15	10	5	10	10	5
restaurant_kitchen	0	0	4	9	9	9	9
shoeshop	5	5	5	5	5	5	5
stairscase	30	25	35	25	35	25	25
studiomusic	37	37	42	21	42	37	42
subway	10	5	5	14	5	5	5
toystore	18	14	18	9	14	18	14
trainstation	35	40	35	40	40	35	40
tv_studio	33	33	22	17	28	28	28
videostore	18	18	18	18	18	14	23
waitingroom	0	0	0	0	0	0	0
warehouse	10	5	14	0	14	14	10
winecellar	14	19	19	19	14	24	24

Figure 4: Experimental results for scene recognition ((b), Amadeus-16 database, and (d-e) Indoor-67 database, results per class and average accuracy) and video retrieval (average precision results on the Semantic Indexing Task of Trecvid 2010 (c-))

Both signatures are used as input for 1-vs-all SVM with chi-square kernel. The Multi-MEDA-based results are then obtained by feeding a Multi-MEDA kernel-based 1-vs-all SVM with the MEDA vector. The three contributions of MEDA, BoW and Multi-MEDA are then combined with weighted linear fusion.

As shown in Figure 4 (d-e), the categorization framework proposed for this database benefits from the introduction of the Multi-MEDA kernel. Not only the MultiMEDA brings an improvement of about 6% over the BoW model, but also its combination with MEDA brings a further improvement over the BoW-only model. Moreover, the fusion of Multi-MEDA with the other models is very effective in terms of average accuracy: the combination of the three approaches brings an improvement of 20% compared to the BoW-only classification.

6.2 Video Retrieval

We use the TrecVid 2010 dataset to test the effectiveness of our proposed approach in a video retrieval task. In particular, we focus on the challenging Light Semantic Indexing Task (SIN), of TrecVid [14] 2010, where the participants are asked to build systems that rank relevant shots according to their pertinence to a given set of semantic concepts. The system is composed of a set of features and a set of concept specific SVMs that learn how to distinguish between relevant and non relevant shots based on the distribution of the input signatures. For a new shot, each classifier gives a concept score, and concept scores from different features are linearly combined to obtain the final score for each shot. In our framework, we extract 128-length SIFT features extracted from interest points based on Harris point detector. From such points we extract the following signatures:

- BoW with 500 words
- MEDA with a number of bin per dimension that have been adapted for each concept (typically 10)

BoW and MEDA are learnt using a chi-square kernel. The Multi-MEDA kernel is then applied on top of the MEDA signatures and results are compared with Mean Average Precision. Results in Fig. 4 (a-c) shows that the kernelized solution that we propose in this paper is a good source of information for CBMR. Multi-MEDA, as a stand-alone model, brings an improvement of around 13% to both traditional MEDA and BoW models. The concepts for which MEDA was not performing as good as BoW (e.g. Bus, Telephones, Airplane_Flying) benefit from the multidimensional modeling in the learning phase. In the TrecVid results we can also clearly notice the complementarity of the kernelized multidimensional modeling that we propose in this paper with respect to the exsiting approaches. As a matter of fact, the combination of just two out of the three models (e.g. MEDA + Multi-MEDA) considered for this task gives an average improvement of 30% compared to using the traditional BoW model only. Moreover, when we fuse the contribution of MEDA, Multi-MEDA and BoW together we obtain a prediction on the test set that is 50% more precise compared to traditional aggregators alone.

7. CONCLUSIONS, LIMITATION AND FUTURE WORK

We proposed a kernel function for MEDA signatures. Multi-MEDA performs at a kernel level a multivariate analysis of the feature space given the marginal approximations stored in the MEDA vector. The resulting model is a multidimensional representation of the LED space, built without explicitely defining a multidimensional codebook nor a new, complex image signature. We showed that by embedding the marginal products into a shift-invariant kernel, the cost of computing such multidimensional model becomes polynomial with the number of dimension of the locally extracted descriptor.By doing so, we allow to model the LED space under a new, more discriminative, and complementary point of view compared to traditional kernels for MEDA signatures. Experimental results show indeed that our new kernel improves the MEDA discriminative power when embedded in a categorization/retrieval framework, and it increases the final retrieval performances when we combine its contribution with traditional kernels over LEDs aggregators (+ 50% on the TRECVID data).

One limitation of our model is the assumption of independence between the LED components. We plan to develop the Multi-MEDA kernel so that it can built a more accurate model that can reconstruct the joint contribution of the LED components by considering the real joint contribution of the marginals. Moreover, Multi-MEDA considers the relationships between the components over k-d bins only. However, we could explore the possibility of consider the joint contributions of the marginals over bins of dimensions $l < k$ (e.g. considering the interactions between components 2 by 2, 3 by 3 etc.), building a complete model with various level of multivariate analysis.

8. REFERENCES

[1] L. Bo and C. Sminchisescu. Efficient match kernel between sets of features for visual recognition. *Advances in Neural Information Processing Systems*, 2(3), 2009.

[2] 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.

[3] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. Ieee, 2005.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–264. IEEE, 2003.

[5] P. Gehler and S. Nowozin. Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2836–2843. IEEE, 2009.

[6] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. 2004.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. Ieee, 2006.

[8] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1):259–289, 2008.

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

[10] 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, pages 2161–2168. Ieee, 2006.

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

[12] M. Redi and B. Merialdo. Marginal-based visual alphabets for local image descriptors aggregation. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 1429–1432, New York, NY, USA, 2011. ACM.

[13] M. Redi and B. Merialdo. A multimedia retrieval framework based on automatic graded relevance judgments. *Advances in Multimedia Modeling*, pages 300–311, 2012.

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

[15] T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *11th IEEE International Conference on Computer Vision (ICCV '07)*, pages 1–8, Rio de Janeiro, Brazil, 2007. IEEE Computer Society.

[16] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. *Computer Vision–ECCV 2008*, pages 696–709, 2008.

[17] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1800–1807. IEEE, 2005.

[18] J. Wu and J. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 630–637. IEEE, 2009.

[19] L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2, pages 721–724. IEEE, 2001.