Direct Modeling of Image Keypoints Distribution through Copula-based Image Signatures

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

Local Image Descriptors (LID) aggregation models such as Bag of Words and Fisher Vectors represent an image based on the distribution of its LIDs *given* a global model, e.g. a visual codebook or a Gaussian Mixture.

Inspired by Copula theory, in this paper we propose a LID-based feature that represents *directly* the behavior of the image LID distribution, without requiring to compute a global model. Following the definition of Copula, we represent the distribution of the image LIDs by describing, on one side, its marginals, and on the other side, a Copula function. The Copula defines the dependencies between the marginals and their mapping to a multivariate probability distribution function. We test the resulting feature for scene recognition and video retrieval (Trecvid data), showing that our approach outperforms, in both tasks, the Bag of Words and the Fisher Vectors Model.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Feature Extraction

Keywords

Scene Recognition, Feature Extraction, CBIR, Gaussian Copulae

1. INTRODUCTION

Content-based image recognition and retrieval (CBIR) techniques are of crucial importance for the management of large collections of multimedia data. CBIR systems build models of the image space by learning image signatures with kernel machines. One of the key elements for the development of effective CBIR systems is the discriminative power of the image signature.

Due to their high discriminative power, signatures based on the *aggregation of local image descriptors* (LIDs) received a lot of attention in the recent years. Among those, the Bag of Words (BoW) model [1] is probably the most widely used method for LID-based analysis. According to this model, k-dimensional LIDs are first extracted from the surrounding of interest [5] or dense [4] points in a set of training images, and then clustered into a visual codebook. Such codebook is then used to map each new image into a fixed length signature, approximating the *multivariate* probability density function (PDF) of the image LIDs given the global codebook. Similarly, Fisher Vectors [8] approximate the distribution of the image LIDs by analyzing, with Fisher Kernels, the similarity between the PDF of the image LIDs and the global PDF of all the LIDs in a training set. Both approaches *represent the joint probability of the image LIDs indirectly*: they describe the behavior of the LIDs in an image *given* a model of the global LIDs space, obtained through operations (generally very expensive) in the k-dimensional space, such as clustering or mixture modeling. Despite its proved accuracy, this type of representation leads to a lack of discriminative power for complex classification tasks, and to high computational complexity in the training phase.

The MEDA [18] signature is an alternative algorithm for LIDs aggregation that partially addresses these problems. In MEDA, each dimension of the LID is quantized into n unidimensional bins. The MEDA feature vector is then built to represent the collection of the k approximated *monovariate* marginal functions. Despite its efficiency, one of the major issues regarding MEDA is that the one-dimensional quanti-

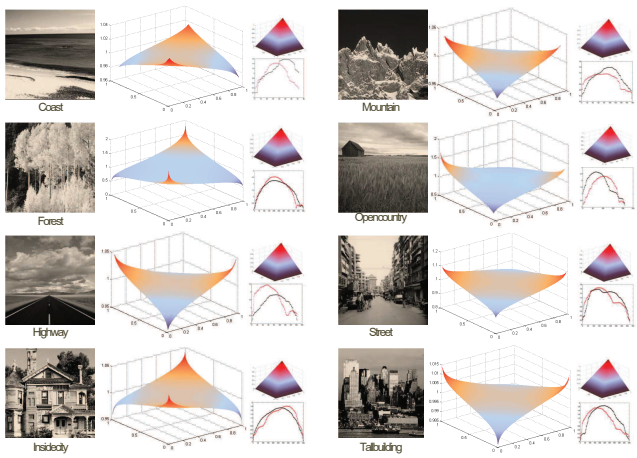


Figure 1: The shape, for different classes, of Gaussian Copula PDF (big plot), CDF and marginals (small plot) arising from the first two dimensions, , i.e. the most informative, of the set of image PCA-SIFT [9].

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.

zation *breaks the correlation between the LIDs components*, losing a lot of precious information arising from the intra-dimension relations and the multivariate LID modeling.

Our idea is to build a LID-based feature vector that can compensate this loss of information. Copula theory [22] tells us that marginals can actually play an important role in multivariate modeling. According to this theory, the PDF of a k-dimensional vector X can be decomposed into k marginal distributions and one Copula function. While the marginals describe the probability of each variable of the random vector, the Copula function represents the dependencies between the marginals, and defines the probability of the vector by mapping the marginal PDF of the variables to their joint PDF. Such mapping is either pre-defined or calculated based on the marginal values, without therefore involving computationally expensive multidimensional searches. For this reason, Copulae are employed as efficient tools for multivariate modeling, and widely adopted in financial and medical data analysis. Here, we apply Copulae to CBIR and LID-based analysis. The main intuition is that, for an image I, we can fit a Copula with the marginals of the LIDs in I, and then describe I according to the resulting PDF shape. Following Copula Theory, in order to build such representation, we should study separately the LIDs marginals and their dependencies.

Given these observations, in this paper we present COMS (COpula and Marginals Signature): a Copula-inspired extension of MEDA that, by using Copulae, allows for *efficient multivariate analysis of image LIDs using pure marginal information*. COMS combines the MEDA vector with its complementary feature, that we name CoMEDA - Copula over MEDA. While MEDA models the pure monovariate information of the marginal *distributions*, CoMEDA represents the Copula structure: the marginal *dependencies*, namely the mapping between the LIDs marginal values and the LIDs joint density. The resulting COMS feature (MEDA + CoMEDA) reflects directly the PDF of the LIDs in an image, without involving the estimation of a global LID model such as visual codebooks. COMS is therefore much more discriminative and much faster in the training phase compared to both Fisher Vectors and BoW.

How do we model such feature? In our approach, we focus on a particular type of Copula, the Gaussian Copula. This function describes the CDF (Cumulative Distribution Function[1]) of a random vector through the shape of a multivariate normal CDF with the following properties: (1) its variables are the normal inverse of the marginals of the vector, (2) its covariance matrix is the correlation matrix between the marginal inverses and (3) its mean is zero. The Gaussian Copula function depends on one parameter only, namely its covariance/correlation matrix, corresponding to the dependencies between the marginals. We therefore store in the CoMEDA vector the values of the correlation coefficients of the marginal inverses. By doing so, we represent in a single feature the marginal dependencies determining the Copula structure. We then match the CoMEDA features using traditional kernel machines such as Support Vector Machines.

Despite the accuracy of CoMEDA as a stand-alone descriptor for CBIR, we know from Copula Theory that we can achieve a complete representation of the image PDF only when we combine marginals and Copula together. We therefore concatenate MEDA and CoMEDA in a single, very discriminative, Copula-inspired image descriptor, COMS, which we then use as input for the learning system.

We test the effectiveness of our approach by comparing it with existing methods in two challenging tasks, namely scene recognition (for small-scale indoor/outdoor scenes [17, 16], and large scale scene recognition on sun database [25]), and video retrieval (TRECVID data [23]). We show that the Copula-based model outperforms traditional BoW-based and Fisher Vectors-based classification.

The remainder of this paper is organized as follows: in Sec.2 we outline the related work in the field. We then explain the statistical differences between our proposed approach and the existing methods in Sec.3, and in Sec.4 we give some highlights on Copula Theory. Sec.5 explains in details our approach and finally Sec.6 validates our theory with experimental results.

2. RELATED WORK

In this section we outline the research works that directly link with our approach. Since the CoMEDA feature is based on local image descriptor aggregation, and inspired by Copulae, we will here first summarize the relevant work concerning LID-based image representation, and then highlight the related work from Copula Theory.

Features based on LID aggregation can be divided into two groups, based on the type of LID probability distribution they are trying to approximate: *multivariate* and *monovariate* approaches.

Multivariate LIDs Aggregators. As mentioned in the previous section, the BoW model is probably the most popular framework for image representation based on locally extracted descriptors. It aims at describing the image based on the LIDs global density, by vector quantizing the LIDs space into a set of visual codewords. The BoW was first introduced by Csurka et al. in [1], applying k-means clustering on a training set of LIDs and then using the centroids of the resulting clusters as visual words. Various techniques have been proposed later on to vector quantize the LID space and improve the construction of the visual codebooks. For example, in [11] mean-shift clustering is used, [15] hierarchically quantizes LIDs in a vocabulary tree and [13] uses Extremely Randomized Clustering Forests to build efficient visual codebooks. Another way to define visual codebooks is proposed in [24], where the codebook is composed of the hypercubes resulting from the quantization of each dimension of the LID into a fixed lattice. While generally the visual word assignment is performed by counting the number of occurrences of each visual word in a given image, Jegou et al in [7] improve this approach by computing, for each point, the element-by-element distance with the closest visual word, and store in the VLAD vector the resulting values. Similar to this approach, Perronnin et al. in [8] first estimate the global LIDs density using Gaussian Mixtures over a LID training set, and then use Fisher Kernels over image keypoints to generate the Fisher Vector signatures, that reflects the way in which the parameters of the image LIDs distribution should be changed to fit the global Gaussian Mixture. Fisher Vectors are proved to be one of the most effective solutions for LID-based image analysis. Despite its computational cost, the multivariate analysis performed by

[1]As we will see later, the Copula-based PDF is easily inferable from the equation of a Copula CDF

the mentioned approaches leads to quite accurate features for CBIR.

Monovariate LIDs Aggregators. In order to overcome some of the computational issues, and to highlight the discriminative power of the LIDs marginals, a 1-dimensional search approach was proposed in [18]. The MEDA descriptor in [18] concatenates the marginal approximation by counting the occurrences of the LIDs component on a predefined set of one-dimensional bins. This approach is very efficient, and it provides a new source of information in the LID analysis, because it exploits the marginal information. However, such monovariate analysis breaks the relations between the LID components, losing precious information for image discrimination. The MultiMEDA kernel presented in [19] is a first attempt to improve the MEDA analysis by adding some multivariate information: assuming component independence, MultiMEDA multiplies the LID marginal values, generating a multidimensional probability out of the MEDA marginal approximations.

Even if MultiMEDA improves the MEDA discriminative power, it is still based on the assumption that the LIDs components are independent and that their marginals are uncorrelated. However, LID vectors arise from the analysis of an entire image region, and each element in a LID is crucial to define the surroundings of an interest point. It is therefore important to analyze the real multivariate information that characterizes those vectors. Given these observations, our idea is to use Copulae to build a complete multivariate analysis of the LID space and generate a feature vector out of such analysis. Why Copulae? Copulae are statistical tools for linking the marginals of the variables in a random vector with their multivariate joint distribution, modeling separately marginal distributions and their dependence structure. We can therefore use them to analyze the LIDs multivariate density by using marginal distributions only, in an efficient and statistically meaningful way.

Copulae first appeared in [22] in the field of probabilistic metric spaces, and they were then widely adopted in finance and actuarial sciences. In particular, Gaussian Copulae are very popular in civil engineering and medical computations, due to their efficiency for multivariate modeling. They allow estimating the joint probability of a vector in a quadratic time, overcoming many computational problems of multivariate modeling. Copulae have indeed been employed in literature for clustering on simulated data [3], and on scientific data [2]. In the image processing domain, Copulae have been used for vector quantization in image coding [6] and for dual polarization synthetic aperture radar image analysis [10]. The only work that applies Copulae to CBIR is, to our knowledge, the work in [20] for the construction of efficient visual codebooks . COMS is, as far as we know, one of the first attempts to build a LID-based feature vector for CBIR using Copulae over the image LIDs.

Overall, our approach is different from all the mentioned approaches because we are not analyzing the independent marginal behavior (such as [18] and [19]), but we are instead trying to estimate the multivariate density of the image LIDs through Copulae. However, we do not compute any global model through clustering [1, 15, 11], Gaussian mixture modeling [8], or other operation in the k-dimensional space [24]. We directly estimate the PDF of the image LIDs and then store in COMS the parameters of such distribution. COMS is therefore statistically different from the space determined

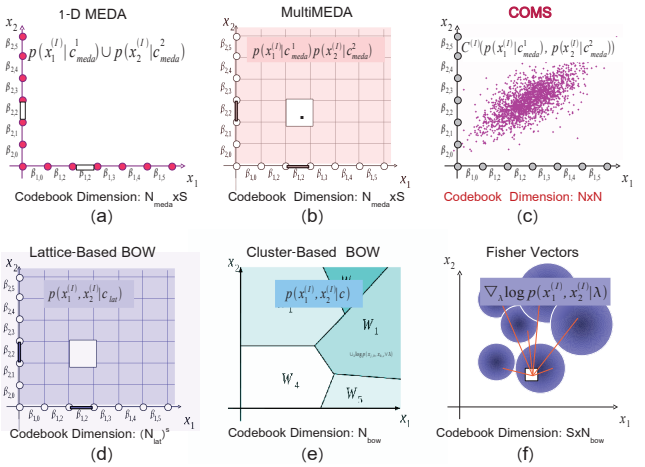


Figure 2: Comparison between the existing LID aggregators and our Copula-Based approach, based on the probabilistic analysis they perform on the image LIDs.

by BoW, MEDA and Fisher Vectors. We will see in the next section a detailed analysis of those differences.

3. WHAT ARE WE MODELING?

In this section, we will show the novelty introduced by the COMS with respect to existing approaches, and show that it represents a new source of information about the LIDs space. Assume for an image I we have a set of m k-dimensional LIDs $x^{(I)} = \{x_j^i\}_{j=1,\dots,k}^{i=1,\dots,m}$: in the following, we will outline the type of probabilistic analysis performed on the LIDs x by the most popular methods for LID-based image analysis.

Multivariate modeling approaches. In the **BoW-like models** [1, 15, 11] a codebook c of N_{bow} k-dimensional vectors is obtained through the clustering of the LID of a training set. An approximation of the joint probability of the LIDs $p_{bow}(x^{(I)}) = p(x^{(I)}|c)$ is then obtained by counting the occurrences of the visual words in an image, see Fig. 2 (e). Similarly (Fig. 2 (d)) **Lattice- based BoW** models like [24], build a vocabulary of hypercubes generated through the quantization of each dimension of the LID in a fixed number of N_{lat} bins, and then reduce such vocabulary c_{lat} according to the informativeness of the resulting codewords. **Fisher Vectors** [8] approximate the LIDs joint distribution by first estimating a universal Gaussian Mixture Model (GMM) on a training set, as shown in Fig. 2 (f). They then compute the log likelihood of the image LIDs with respect to the parameters λ of the GMM. Finally, they store in a feature vector the concatenation of the resulting partial derivatives, namely $p_{fv}(x^{(I)}) = \bigtriangledown_\lambda \log p(x^{(I)}|\lambda)$.

Monovariate modeling approaches. Opposite to the other approaches, the **MEDA** [18] model generates a codebook c^j_{meda} of $1-d$ letters through one-dimensional marginal quantization, see Fig. 2 (a). The resulting MEDA vector represents the approximation of the marginals: $p_{meda}(x^{(I)}) = \cup_{j=1}^{k} p_j(x_j^{(I)}|c^j_{meda})$. An extension of MEDA that allows for multidimensional probability estimation given marginals is **Multi-MEDA** [19] (Fig. 2 (b)), that performs a kernelized Cartesian product of the marginal approximations in MEDA, assuming independence between LIDs components, giving $p_{Mmeda}(x^{(I)}) = \prod_{j=1}^{k} p_j(x_j^{(I)}|c^j_{meda})$.

*Our Approach.*Our approach, **COMS**, depicted in Fig. 2 (c), is different from all the other approaches, and lies in an intermediate point between the marginal and the multivariate analysis. We estimate the joint distribution of the image through a Gaussian Copula $C_\Sigma(p_1, \ldots, p_k)$ over the MEDA-based approximated marginal distributions. This leads to a reliable representation of the multivariate LIDs distribution given the image monovariate marginal approximations. The peculiarity of the Copula-based distribution is that it depends on one parameter only, namely the correlation Σ between the inverse of the image marginals. We therefore first store in CoMEDA the values of Σ directly, giving $p_{CoMeda}(x^{(I)}) = corr(p_1^{-1}(x_1^{(I)}|c_{meda}^1), \ldots, p_k^{-1}(x_k^{(I)}|c_{meda}^k))$. CoMEDA represents the multivariate complement of the monovariate MEDA vector, since it represents the dependencies between the marginal distributions. Therefore in COMS, MEDA+CoMEDA, namely the union of the two fundamental element of the LIDs density according to Copula theory, we model a complete Copula-based distribution, $p_{COMS}(x^{(I)}) = C_\Sigma(p_1(x_1^{(I)}|c_{meda}^1), \ldots, p_k(x_k^{(I)}|c_{meda}^k))$.

The Copula modeling that we perform is therefore different ($p_{COMS} \neq p_{meda}, p_{Mmeda}$) from the marginal modeling approaches, because, despite the underlying marginal analysis, it does not assume independence between the LID components, but it models instead a real joint PDF based on the marginal dependencies. On the other hand, we can also say that $p_{COMS} \neq p_{bow}, p_{fv}$ because, first of all, the shape of Copula-Based joint probability described by COMS is different from the shape of the Mixture Model estimated by BoW and Fisher Vectors. This suggests that, by introducing COMS in the LIDs-based analysis, we add some new, complementary information regarding the LIDs distribution. Moreover, due to the simplicity of the Copula algorithm, we can build one Copula per image representing its joint PDF: we can then discriminate between different images using the distribution information given by the specific shape of the image Copula. While BoW approximates the joint LIDs distribution through Vector Quantization given a global codebook, and while Fisher Vectors store the results of parameter adaptation for GMM fitting, COMS directly stores the parameter of the image joint PDF, leading to a more informative image feature. Since both MEDA and CoMEDA arise from the analysis of the image LIDs marginals, COMS does not require an unsupervised search on a training set in the k-dimensional space such as GMM, k-means or hypercube exploration to define the global LID density, saving a lot of computational time on training.

4. COPULA THEORY

Copulae [22] are structures that allow linking the marginal distributions of the variables in a random vector with their joint density function. While traditional multivariate analysis combines the study of marginal and joint densities, Copula Theory provides statistical models to study separately the marginal distributions and their dependencies. The main idea is that the joint distribution of the variables in a random vector X of length k can be decomposed into k marginal distributions and a Copula function C. C represents the link between the marginals: their dependency structure, their mapping to a multivariate cumulative distribution function (CDF).

Such mapping is either given explicitly or, as in our case, inferred through the analysis of the marginal behavior, without recurring to complex multivariate modeling. The advantage of using Copulae for multivariate modeling is therefore that they can estimate the multidimensional distribution of a random vector very efficiently, given just the values of its marginals. In this section, we give some highlights on the Copula theory, and in particular on Gaussian Copulae, that we will then apply for CBIR purposes. It is outside of the scope of this paper to cover all the aspects of Copulae, therefore we will introduce only the basic tools to understand our COMS feature. For ease of understanding, we will outline the theory through the bivariate case analysis ($k = 2$), that is easily extendable to the multivariate scenario.

4.1 Copulae: Linking Marginals with Joint Distributions

Given a 2-dimensional random vector $x = \{x_1, x_2\}$, we define $u = F_1(x_1) = [P(x_1 \leq X_1)]$, $v = F_2(x_2) = [P(x_2 \leq X_2)]$ as the marginal cumulative distribution functions (CDFs) of X_1 and X_2 respectively, and $F(x_1, x_2) = P[x_1 \leq X_1, x_2 \leq X_2]$ as the vector cumulative joint distribution.

As said, a Copula C, is defined as a unique mapping[2] that assigns the joint CDF of X given each ordered pair of values of its marginals, namely:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)) = C(u, v),$$

and, following Sklar's theorem and assuming that F_1, F_2 are continuous:

$$C(u, v) = F(x_1, x_2) = F(F_1^{-1}(u), F_2^{-1}(v)), \qquad (1)$$

which allows to construct a Copula from a given multivariate distribution function F.

The Copula function by itself describes the vector CDF. However, we might want to represent the vector in terms of probability density function (PDF), i.e. $f(x_1, x_2) = P[x_1 = X_1, x_2 = X_2]$. In order to obtain $f(x_1, x_2)$ we have to compute *copula density*, namely the CDF derivative, i.e., following Eq. (1) :

$$f(x_1, x_2) = \frac{\delta^2 C(u, v)}{\delta u, \delta v} = \frac{f(F^{-1}(u), F^{-1}(v))}{f(F^{-1}(u)), f(F^{-1}(v))},$$

where f is the PDF corresponding to F.

The copula describes therefore the dependence between the components of a random vector, no matter the function describing their marginal distributions: if we know the mapping C, the joint distribution $f(x_1, x_2)$ can be inferred from the marginal CDFs u and v.

4.2 Gaussian Copulae

A particular type of Copulae is the Gaussian Copula, which belongs to the class of Elliptical Copulae (i.e. Copulae following Elliptical distributions such as Laplacian, T-Student, etc..). The Gaussian Copula structure is a multivariate normal distribution: in this model, F corresponds to the multivariate Gaussian CDF, while F^{-1} corresponds to the inverse of the univariate normal CDF.

[2]In order to be defined as a two-dimensional Copula, C needs to fulfill the following requirements (see [14]):

- It is defined over the interval $[0, 1]$
- $\forall t \in [0, 1]$, then $C(t, 0) = C(0, t) = 0$ and $C(t, 1) = C(1, t) = 1$
- $\forall u_1, u_2, v_1, v_2 \in [0, 1]$, with $u_1 \leq u_2$ and $v_1 \leq v_2$, $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$

A Gaussian Copula C_Σ is then defined for the 2-dimensional random vector x as (following Eq. (1)):

$$C_\Sigma(u,v) = \phi_\Sigma(\Phi^{-1}(u), \Phi^{-1}(v)), \tag{2}$$

being $\Phi^{-1}(\cdot)$ the inverse of the univariate normal CDF, and ϕ_Σ the bivariate (or multivariate, when $k > 2$) standard with mean zero and covariance Σ, , giving

$$C_\Sigma = \frac{1}{\sqrt{det(\Sigma)}} exp\left(-\frac{1}{2} \cdot \begin{pmatrix} \Phi^{-1}(u) \\ \Phi^{-1}(v) \end{pmatrix}^T \Sigma^{-1} \mathbf{I} \begin{pmatrix} \Phi^{-1}(u) \\ \Phi^{-1}(v) \end{pmatrix} \right), \tag{3}$$

How to find the covariance matrix Σ? When dealing with normal distributions, the correlation values between two variables fully define their dependencies. In Gaussian Copulae, Σ corresponds therefore to the correlation matrix between the inverse standard univariate normal CDF

$$\Sigma(\Phi^{-1}(u), \Phi^{-1}(v)) = \frac{cov(\Phi^{-1}(u), \Phi^{-1}(v))}{\sigma(\Phi^{-1}(u))\sigma(\Phi^{-1}(v))} \tag{4}$$

4.3 Why Gaussian Copulae?

As said, the Gaussian Copula function arises from pure marginal analysis: both the variables (inverse normal of marginal CDFs) and the parameter (correlation between the inverse marginal CDFs in Eq. (4) are constructed by manipulating the marginal distributions with simple operations ($O(k)$ for $\Phi^{-1}(\cdot)$, and $O(k^2)$ for Σ). Gaussian Copulae represent therefore an efficient way to estimate the joint PDF of vectors that (I) have a small dimensionality, namely a low value of k and (II) have marginals that can be easily modeled. In fact, local image descriptors satisfy conditions (I) and (II). The dimensionality of LIDs is generally $k \leq 128$. Moreover, it exists a descriptor for LID marginal approximation, MEDA, which have been proved to effectively model the univariate distributions of the LID components. Gaussian Copulae can be therefore very efficient tools to estimate the joint PDF of LIDs.

Moreover, a Gaussian Copula C_Σ depends on one parameter only, namely the covariance-correlation matrix Σ, whose computational time that is quadratic with k, making it easy to characterize an image through its Copula shape. Furthermore, various fast implementations are available to easily and fastly treat with multivariate normal densities, due to their popularity, making the computation of this Copula very easy. This motivates us to use Gaussian Copulae to efficiently and effectively approximate the distributions of the LIDs in an image and generate an image signature out of it.

5. COMS: MULTIVARIATE LID ANALYSIS FROM MARGINAL VALUES

In this section, we show how to exploit Copulae Theory to aggregate LIDs and build effective and efficient compact image signatures based on local descriptors.

In order to perform LID-based analysis, for each image I, we first extract m salient points and describe them using a k-dimensional normalized SIFT [12] descriptor $x^{(I)} = (x_1^i, \ldots, x_k^i)$, $i = 1, \ldots, m$. For an image I, we define $p_j(x_j^{(I)})$ and $P_j(x_j^{(I)})$ $j = 1, \ldots, k$, as the marginal probability distribution and cumulative distribution of the j_{th} component of the image LIDs, and $p(x^{(I)})$ as their joint density.

The main idea is that, similar to Copula Theory, we can approximate $p(x^{(I)})$ for an image I by extracting (A) its set of marginals $p_j(x_j^{(I)})$ and (B) a Gaussian Copula Function, and use it as a discriminative image signature for CBIR purposes. While (A) it already exists a feature (i.e. MEDA) approximating the marginals, we are missing (B) a feature to represent the Copula structure. We therefore design CoMEDA for this purpose (See Fig. 3 for a visual explanation of our approach).

Therefore, we first (A) extract from image I the MEDA vector $v^{(I)}$ containing the LIDs marginals approximations. We then (B) use them, as shown in Sec.5.2, to estimate the marginal CDFs and fit an image-specific Gaussian Copula $C_\Sigma^{(I)}$, that defines an approximation of the joint distribution of the image LIDs. We characterize the image I with the Copula structure of its LIDs by storing in the CoMEDA feature the values of the image-specific covariance matrix $\Sigma^{(I)}$, namely the unique parameter of the resulting Copula-based PDF. Finally, we achieve a complete model of the LID density by combining the CoMEDA feature of an image I with its marginal counterpart, i.e. the MEDA vector for image I, into a final image signature, namely COMS.

5.1 MEDA: Modeling Marginal Distributions of Local Features

In order to build a complete Copula-based representation of the Image I, we first perform marginal analysis through the MEDA descriptor. The MEDA descriptors [18] were designed to highlight the discriminative power of the LIDs marginals. The MEDA signature represents the concatenations of the approximations of the k marginals of the image LIDs.

First, it quantizes each component j of the LIDs in an image I, in a set of n discrete bins $\beta_{j,b}$, $b = 1, \ldots, n$. The MEDA vector is then produced by collecting the frequencies of such bins over the set of x_i extracted from an image I. By doing so, MEDA describes the univariate behavior of the image LIDs, and stores in a single descriptors the set of k approximated marginals distributions $\tilde{p_j}(x_j^{(I)})$. As a matter of fact, the final image signature $v^{(I)}$ is a $k \times n$ histogram, obtained by counting how many LIDs at a given dimension j fall into a given bin b

$$v^{(I)}(j,b) = p(x_j|\beta_{j,b}) = \#\{x^i : x_j^i \in \beta_{j,b}\}.$$

5.2 Fitting a Copula with the Image LIDs

Once we have extracted the marginal information from the Image LIDs, we should calculate the corresponding Gaussian Copula. This will allow us to characterize each image with the distribution of its LIDs (using the parameters of the Copula-based density as signature).

First, for each dimension of the LID, for each of the k marginals $\tilde{p_j}(x_j^{(I)})$ that we obtain with the MEDA histogramming[3], we compute the corresponding k univariate CDFs $u^{(I)}(1) = P_1(x_1^{(I)}), \ldots, u^{(I)}(k) = P_k(x_k^{(I)})$, normalized in the interval $[0, 1]$. According to the Gaussian Copula theory, we then compute normal inverse CDF over the resulting LIDs Cumulative Distribution Functions, namely

$$\Phi^{-1}(u^{(I)}(1)), \ldots, \Phi^{-1}(u^{(I)}(k)). \tag{5}$$

[3]In practice, we will use for our experiments a more refined way to estimate the marginal distribution shape, namely a kernel density estimator [21]

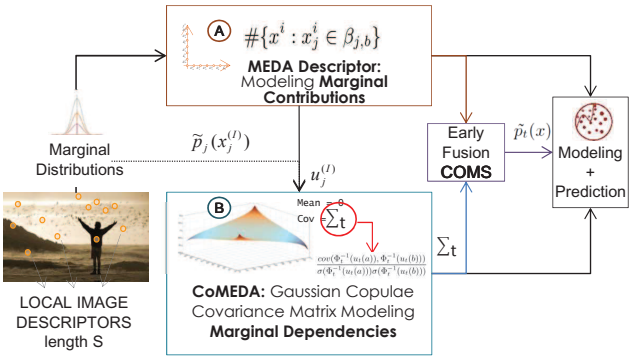


Figure 3: Our Copula-based LID aggregator.

If we now want to define a Gaussian Copula $C_{\Sigma}^{(I)}$ representing the CDF of the LIDs for image I, we should extend the multivariate Gaussian in Eq. (2), for SIFT vector analysis with $k \gg 2$, giving, for image I,

$$C_{\Sigma}^{(I)}(u^{(I)}(1),\ldots,u^{(I)}(k)) = \phi_{\Sigma}^{(I)}(\Phi^{-1}(u^{(I)}(1)),\ldots,\Phi^{-1}(u^{(I)}(k))). \quad (6)$$

and from the Copula theory, we know that $\Sigma^{(I)}$ can be computed as the correlation matrix between the inverse of the LID marginals, namely:

$$\Sigma^{(I)}(a,b) = \frac{cov(\Phi^{-1}(u^{(I)}(a)),\Phi^{-1}(u^{(I)}(b)))}{\sigma(\Phi^{-1}(u^{(I)}(a)))\sigma(\Phi^{-1}(u^{(I)}(b)))} \quad (7)$$

where $a,b = 1,\ldots,j$, $cov(\cdot,*)$ corresponds to the covariance between $(\cdot)$ and $(*)$, and $\sigma(\cdot)$ is the standard deviation of variable $(\cdot)$.

5.3 The CoMEDA vector

How can we capture the behavior of the Copula structure we just described, and store it into a single effective feature? As we can observe, Eq. (6), has only one parameter, the covariance matrix $\Sigma^{(I)}$. Such covariance matrix describes the dependencies structure between the LIDs marginals and determines the equation of the multivariate distribution.

We therefore fill the CoMEDA vector $\mu^{(I)}$ for an Image I with the values corresponding to $\Sigma^{(I)}$, namely the correlation coefficients of the inverse marginal approximations of the LIDs in the image. The complexity of CoMEDA is quadratic with the number of dimensions of the LIDs, and its dimensionality is $\frac{k\times k}{2}$, being $\Sigma^{(I)}$ typically a symmetric matrix. CoMEDA does not imply therefore exponential computation or multidimensional vector quantization for multivariate LID representation. This low dimensional feature (we will select $k = 36$) can be easily then used as input for discriminative classifiers, that will learn a model of the LIDs space based on the CoMEDA feature representation.

5.4 COMS: MEDA + CoMEDA

CoMEDA gathers the main element of the Copula structure: it stores the LIDs multidimensional information arising from the dependencies between marginal distributions.

However, we can observe that the shape of Eq. (2) is determined both by $\Sigma^{(I)}$ and by the behavior of the LIDs marginal distributions, specific of the image I. Recall that, as a matter of fact, Copula theory states that the joint distribution of a random vector can be represented by its marginal distributions and a multivariate Copula structure. This suggests us that, in order to have a complete representation of the LID space, we should combine the CoMEDA feature of image I with a descriptor approximating the marginal behavior of I, e.g. MEDA. Therefore, for each image, we concatenate these two types of information regarding the LID distribution, MEDA and CoMEDA, both very discriminative features, into a single image descriptor COMS $h^{(I)} = \{v^{(I)},\mu^{(I)}\}$. By doing so, we enrich the representation of the LID space, and determine a good approximation of the complete LID joint distribution.

6. EXPERIMENTS

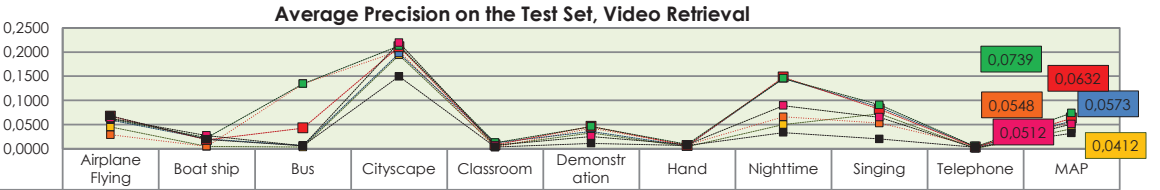
In this section we will show the performances of our Copula-based approach, comparing it with the most effective LID aggregators available in literature. We test the effectiveness of our approach for two, challenging tasks, namely video retrieval and scene recognition.

Since all the descriptors work over the same input, namely local image descriptors, the first step of our experiments is to compute the image LIDs. Since we want to keep the dimensionality low, from all the images/keyframes in our datasets we compute PCA-Sift [9] ($k = 36$) around interest points extracted with the Hessian detector. We then aggregate them using the following approaches for comparison:
(1) *Bow*, the Bag of Words Model computed, as in [1], through a codebook built with k-means clustering
(2) *Meda*, the marginal-based descriptor in [18]
(3) *Fisher*, the Fisher Vectors approach, computed using and adapting the implementation in [8]
(4) *CoMeda*, our Copula-based descriptor, i.e. the values of the correlation coefficients of the inverse of the marginals
(5) *COMS*, the early combination of MEDA and CoMEDA

Moreover, in order to prove the reasonableness of our Copulae-based LID processing, we compute another feature, that we call MVN (Multivariate Normal), that stores the values of the mean and covariance matrix of the image LIDs vectors (different from CoMEDA, that treats LIDs marginals). The difference of effectiveness between COMS (or other multivariate approaches) and MVN will show the discriminative value added by treating the LIDs with models more complex than a simple multivariate Gaussian PDF.

Then, we use the computed descriptors as input to Support Vector Machines (SVM) with chi-square or Radial Basis Function kernels to build models able to predict the image category, or the presence of a given concept (in the case of Video Retrieval). Finally, in order to further prove the effectiveness of the combination of MEDA and CoMEDA, we combine and weigh the predictions coming from the MEDA-only model and the CoMEDA-only model, and we name this class of experiments *Posterior*.

We show that our approach outperforms the other methods in all the databases considered for scene recognition and for video retrieval. Overall, we can say that posterior fusion of MEDA and CoMEDA is slightly more effective than *COMS*, because we add one parameter to weigh the contribution of the two descriptors. We can also observe that the simple MVN descriptor has a weaker discriminative power compared to all the other descriptors, suggesting that adding complexity in the LID modeling actually is useful for CBIR performances improvement. Regarding computational costs, as we can see from Fig. 4 (c), the time to compute CoMEDA, for the training set, has the same order of magnitude as the MEDA feature, because it does not require to estimate a universal model such as the BoW codebook.



	Airplane Flying	Boat ship	Bus	Cityscape	Classroom	Demonstration	Hand	Nighttime	Singing	Telephone	MAP
Bow	0,0448	0,0047	0,0034	0,1936	0,0061	0,0328	0,0044	0,0497	0,0720	0,0004	0,0412
MEDA_36	0,0288	0,0050	0,1341	0,2024	0,0029	0,0425	0,0079	0,0658	0,0530	0,0059	0,0548
Co-Meda	0,0595	0,0206	0,0065	0,1975	0,0124	0,0355	0,0061	0,1455	0,0862	0,0034	0,0573
COMS	0,0672	0,0185	0,0428	0,2108	0,0079	0,0461	0,0063	0,1479	0,0813	0,0030	0,0632
Posterior	0,0617	0,0225	0,1341	0,2129	0,0124	0,0457	0,0089	0,1455	0,0906	0,0044	0,0739
Fisher	0,0632	0,0272	0,0058	0,2194	0,0070	0,0255	0,0087	0,0890	0,0649	0,0012	0,0512
MVN	0,0679	0,0192	0,0062	0,1493	0,0038	0,0112	0,0064	0,0335	0,0202	0,0028	0,0321

(a)

Accuracy on the Test Set, Scene Recognition

	Bow	Meda	CoMeda	COMS	Posterior	Fisher	MVN
Outdoor	71,4892	71,6471	71,6244	**74,951**	**76,2299**	74,0665	69,6271
Indoor	22,8536	22,9252	24,2833	**26,585**	**27,3495**	25,5934	22,245
SUN	8,0554	8,5945	10,0302	**11,234**	**11,4408**	11,1285	8,5340

(b)

Training Set Processing Time (seconds)

	Bow	Fisher	CoMeda	Meda
Outdoor	14429	7031	1023	401
Indoor	86685	38294	9428	87

(c)

Figure 4: Experimental Results for Scene Recognition and Video Retrieval.

6.1 Scene Recognition

In this section we present the results of our experiments for small scale (indoor/outdoor) and large scale scene recognition. The goal for this task is to build a model able to classify test images with the correct class, selected out of a set of pre-defined mutually exclusive categories. We achieve this goal by learning our features with a one-vs.-all multiclass SVM, and assigning the image category according to the classifier that outputs the highest score. The typical evaluation measure for this task is the average accuracy on the test set. In the following we will see the experimental setup and results for the various datasets considered. A visual representation of the results can be found in Fig. 4(b).

6.1.1 *Small Scale Scene Recognition*

Outdoor Scenes: The Outdoor Scenes Dataset [16] contains 2600 color images from 8 categories of natural outdoor scenes. As in [16], we retain 100 images per class for training and the rest for testing. The LID aggregators that we compare are the following: *bow* with 720 visual words; *Meda* with percentile quantization, as proposed in [18], with 10 bins per dimension (it is therefore 360-dimensional), *Fisher* with 64 Gaussians in the mixture (final dimension is 2304), then *CoMeda* (dimensionality 1296), and *COMS*, with 1656 components, and finally *MVN* with 36*36(covariance)+36(mean)=1332 dimensions.
Our results show that, even if *CoMeda* by itself does not outperform *Meda*, when they are combined together with early (*COMS*) and *Posterior* fusion, namely when we follow the Copula Theory approach, the resulting model is much more efficient than both Bag of Words and Fisher Vectors.

Indoor Scenes: The Indoor Scenes Dataset [16] contains around 15000 color images from 67 categories of diverse indoor scenes. Following the approach in [17], we retain 20 images per class for testing and we train our models with the remaining images. The details of the features that we compare follow: *bow* with 1300 visual words; *Meda* with percentile quantization, as proposed in [18], with 10 bins per dimension (resulting in a feature with 288 components), *Fisher* with 32 Gaussians in the mixture (final dimension is 1152), then *CoMeda* (dimensionality 1296), and *COMS*, with 1656 components, and finally *MVN* with 36*36(covariance)+36(mean)=1332 dimensions.

Results for indoor scenes show a similar trend as the experiments on the outdoor scenes datasets. The CoMEDA feature used as a stand-alone descriptor is actually more performing (+6%) than BoW, and it is improved by its combination with the MEDA descriptor (+ 16% of *COMS* and +20% of *Posterior* over *bow*), with a great improvement, 6% over the Fisher Vectors-based classification.

6.1.2 *Large Scale Scene Recognition*

The sun database [25] contains around 899 categories for more than 130, 000 images. As in [25], we select a subset of images spanning 397 scenes consisting in 10 folds that contains, for each category, 50 images for test and 50 for training. The LIDs aggregators that we compute for this database are as follows: *bow* with 500 visual words; *Meda* with uniform quantization, as proposed in [18], with 10 bins per dimension (resulting in a feature with 360 components), *Fisher* with 32 Gaussians in the mixture (final dimension is 1152), then *CoMeda* (dimensionality 1296), and *COMS*, with 1584 components, and finally *MVN* with 36*36(covariance)+36(mean)=1332 dimensions.

In the results for this dataset, we can see a homogeneous accuracy score obtained the COMS/*Posterior*/*Fisher* descriptor, all outperforming by around 40% the simpler approaches such as MEDA and BoW.

6.2 Video Retrieval

For this task, we focus on the challenging TRECVID 2010 [23] light Semantic Indexing Task, where 10 concepts have to be detected in a video corpus of around 400 hours. We use around 60000 shots for training and an equal number for testing. Once we have created the model using SVMs, the test videos are ranked according to their prediction concept score, and results are compared in terms of Mean Average Precision (MAP). Here, we compute MEDA with fixed quantization (with a number of bins tuned, as in [18], for each concept), *bow* with 500 visual words, *Fisher* with 32 Gaussians in the mixture (final dimension is 2304), then *CoMeda* (dimensionality 1296), and *COMS*, with 1584 components,

and finally *MVN* with 36*36(covariance)+36(mean)=1332 dimensions.

As shown in Fig. 4 (a), the effectiveness of our method is even more clear for this challenging task: while *COMS* outperforms *bow* by more than 50% and *Fisher* by 23%, the posterior fusion of MEDA and CoMEDA is further improving the performances of our proposed method for video retrieval, with an increase of around 78 % over BoW and 44% over the Fisher Vector-based retrieval.

7. CONCLUSIONS

We presented a new method for LIDs aggregation. We are inspired from the Copula theory: we exploit the MEDA marginal approximations to feed a Gaussian Copula and build an image signature representing the multivariate PDF of the image LIDs that we name COMS. The resulting image representation is shown to be more discriminative than BoW and Fisher Vectors for image and video retrieval.

The work in this paper can be extended by finding more effective kernels for Copula-based signature matching, such as kernels based on Bhattacharyya distance or Kullback-Leibler divergence. Moreover, we could use different Copula structures, such as Clayton or T-student Copulae and build more discriminative features out of them.

8. 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] E. Cuvelier and M. Noirhomme-Fraiture. Clayton copula and mixture decomposition. *ASMDA 2005*, pages 699–708, 2005.

[3] F. Di Lascio and S. Giannerini. A new copula–based clustering algorithm.

[4] 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.

[5] 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.

[6] X. Guo, L. Wang, J. Zeng, and X. Zhang. Vq codebook design algorithm based on copula estimation of distribution algorithm. In *2011 First International Conference on Robot, Vision and Signal Processing*, pages 178–181. IEEE, 2011.

[7] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[8] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

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

[10] V. Krylov, G. Moser, S. Serpico, and J. Zerubia. Supervised high-resolution dual-polarization sar image classification by finite mixtures and copulas. *Selected Topics in Signal Processing, IEEE Journal of*, 5(3):554–566, 2011.

[11] 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.

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

[13] F. Moosmann, B. Triggs, F. Jurie, et al. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems 19*, pages 985–992, 2007.

[14] R. Nelsen. *An introduction to copulas*. Springer Verlag, 2006.

[15] 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.

[16] 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.

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

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

[19] M. Redi and B. Merialdo. Exploring two spaces with one feature: kernelized multidimensional modeling of visual alphabets. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 20. ACM, 2012.

[20] M. Redi and B. Merialdo. Fitting gaussian copulae for efficient visual codebooks generation. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pages 1–6. IEEE, 2012.

[21] B. Silverman. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986.

[22] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8(1):11, 1959.

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

[24] 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.

[25] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492, 2010.