Local Log-Euclidean Multivariate Gaussian Descriptor and Its Application to Image Classification

Peihua Li, *Member, IEEE,* Qilong Wang, *Member, IEEE,* Hui Zeng, Lei Zhang, *Senior Member, IEEE*

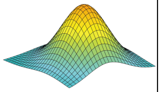
Abstract—This paper presents a novel image descriptor to effectively characterize the local, high-order image statistics. Our work is inspired by the Diffusion Tensor Imaging and the structure tensor method (or covariance descriptor), and motivated by popular distribution-based descriptors such as SIFT and HoG. Our idea is to associate one pixel with a multivariate Gaussian distribution estimated in the neighborhood. The challenge lies in that the space of Gaussians is not a linear space but a Riemannian manifold. We show, for the first time to our knowledge, that the space of Gaussians can be equipped with a Lie group structure by defining a multiplication operation on this manifold, and that it is isomorphic to a subgroup of the upper triangular matrix group. Furthermore, we propose methods to embed this matrix group in the linear space, which enables us to handle Gaussians with Euclidean operations rather than complicated Riemannian operations. The resulting descriptor, called Local Log-Euclidean Multivariate Gaussian (L^{2}EMG) descriptor, works well with low-dimensional and high-dimensional raw features. Moreover, our descriptor is a continuous function of features without quantization, which can model the first- and second-order statistics. Extensive experiments were conducted to evaluate thoroughly L^{2}EMG, and the results showed that L^{2}EMG is very competitive with state-of-the-art descriptors in image classification.

Index Terms—Image descriptors, space of Gaussians, Lie group, image classification.

1 Introduction

Characterizing local image properties has been attracting great research interests in past years [1]. The local descriptors can be either sampled sparsely for describing region of interest localized by region detectors [2], or extracted at dense, regular grids for image representation [3], [4]. Image local descriptors play a fundamental role for the success of many middle-level or high-level vision tasks. In particular, densely sampled descriptors have proven to achieve state-of-the-art performance in image-based classification tasks such as scene categorization, object classification, texture recognition, and image retrieval. However, it is challenging to develop image descriptors with high distinctiveness for general image classification tasks.

TABLE 1
Analogy of Matrix-valued and Function-valued image representation

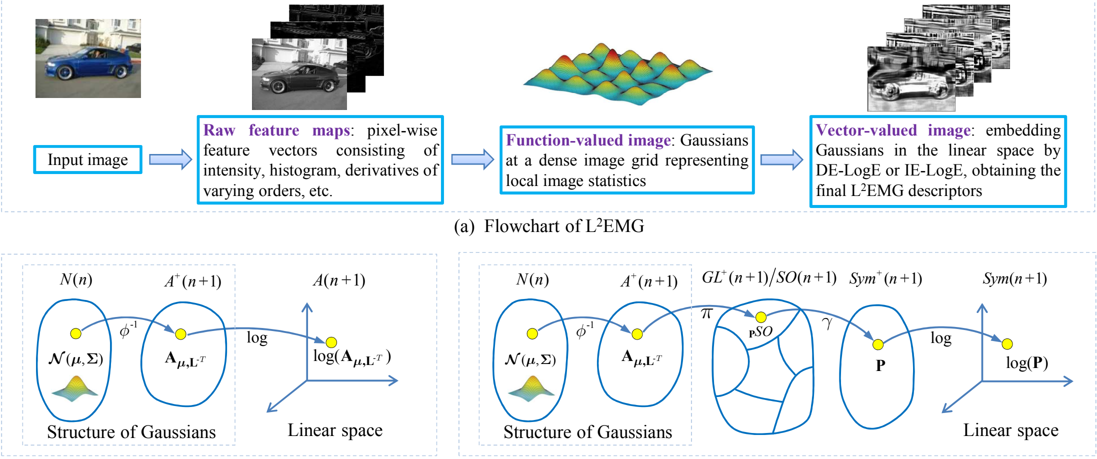
	Representation	Description
STM or COV	$\begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}$	Second-order moment of features which reflects the correlation in local patches.
DTI	$\begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{12} & D_{22} & D_{23} \\ D_{13} & D_{23} & D_{33} \end{bmatrix}$	Every voxel is associated with a 3×3 symmetric matrix describing molecules diffusion.
Ours		We represent the local image statistics using multivariate Gaussians at a dense grid mapped to the linear space via information geometry.

Our goal is to present a function-valued descriptor to effectively represent the statistics of an image local region. Our work is motivated by Diffusion Tensor Imaging (DTI) [5] and the structure tensor model (STM) [6] or covariance descriptors (COV) [7], which respectively enjoy important applications in vision and medical imaging fields [8]. As shown in Table 1, STM or COV computes the second-order moment of image gradients or multiple cues in a neighborhood, which reflects the correlation of features in local image patches. In DTI each voxel is associated with a 3×3 symmetric matrix describing the molecular mobility along three directions and the correlations among these directions. Our idea is to associate one pixel point with a multivariate Gaussian distribution (hereafter abbreviated as Gaussian for simplicity) to characterize the first- and second-order statistics in the local neighborhood. It extends the STM or COV in that both the first- and second-order moments are utilized; similar to DTI, our method can be interpreted as an "imaging" method, which produces a function-valued image where each point is a Gaussian describing the local feature statistics.

Our work is also inspired by the popular distribution-based descriptors, e.g., SIFT [9] and HoG [10]. Such descriptors usually employ histogram to represent the local image statistics. Though effective in a variety of applications, they only exploit zero-order statistics as only feature occurrences (frequencies) are collected [4] and there lacks a natural mechanism for multiple cue fusion. In addition, discrete histograms often suffer from the quantization prob-

- *P. Li, Q. Wang and H. Zeng are with the School of Information and Communication Engineering, Dalian University of Technology, China. E-mail: peihuali@dlut.edu.cn, qlwang, zenghui118@mail.dlut.edu.cn*
- *L. Zhang is with the Department of Computing, The Hong Kong Polytechnic University. E-mail: cslzhang@comp.polyu.edu.hk*

This work was supported by the National Natural Science Foundation of China (61471082), and the Hong Kong RGC GRF grant (PolyU 5313/13E).


Input image
Raw feature maps: pixel-wise feature vectors consisting of intensity, histogram, derivatives of varying orders, etc.
Function-valued image: Gaussians at a dense image grid representing local image statistics
Vector-valued image: embedding Gaussians in the linear space by DE-LogE or IE-LogE, obtaining the final L^{2}EMG descriptors
$N(n)$
$A^+(n+1)$
$A(n+1)$
ϕ^{-1}
log
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$
$\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$
$\log(\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}})$
Structure of Gaussians
Linear space
$GL^+(n+1)/SO(n+1)$
$Sym^+(n+1)$
$Sym(n+1)$
π
γ
$\mathbf{P}SO$
$\mathbf{P}$
$\log(\mathbf{P})$

(a) Flowchart of L^{2}EMG

(b) DE-LogE

(c) IE-LogE

Fig. 1. Overview of the Local Log-Euclidean Multivariate Gaussian (L^{2}EMG) descriptor. (a) Given an input image, we compute raw features and then extract Gaussians at a dense grid using these features; finally we embed the Gaussians in the linear space to obtain vectorized L^{2}EMG descriptors. We show that the space of n-dimensional Gaussians $N(n)$ can be equipped with a Lie group structure, equivalent to a subgroup, denoted by $A^+(n+1)$, of the upper triangular matrix group. It can be directly embedded in a linear space $A(n+1)$ by the matrix logarithm which is called DE-LogE, or indirectly by IE-LogE which first maps $A^+(n+1)$ into the space $Sym^+(n+1)$ of SPD matrices and then into the linear space $Sym(n+1)$, as shown in (b) and (c), respectively. The embedding processes preserve the algebraic and topological structures. Consequently, we can handle Gaussians with Euclidean operations instead of Riemannian ones. See Section 4 for details.

lem which may bring side effects on vision tasks [11]. These considerations instigate us to model image local statistics using Gaussian, one of the most widely used continuous probability density functions. Underlying our descriptor are the maximum entropy principle [12] and the success of covariance descriptors [7]. The maximum entropy principle states that in the set of trial distributions encoding the precisely stated prior or testable information, the one with maximal entropy is the proper distribution. It is well-known that Gaussian enjoys such a property among the family of distributions with fixed empirical mean vector and covariance matrix. The covariance matrices have proven to be effective descriptors in a variety of applications [7]; beside, the mean vectors have also proven to be important in image classification [4] and image search [13].

The challenge of using Gaussians to model local image statistics lies in that the space $N(n)$ of n-dimensional Gaussians $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix, is not a linear space but a manifold. We study the geometry of $N(n)$, and show, for the first time to our knowledge, that the space of Gaussians can be provided with a Lie group structure by defining a multiplication operation on this manifold, and that it is isomporhic to a subgroup, denoted by $A^+(n+1)$, of the upper triangular matrix group. Based on this, we develop novel methods to embed this space into some linear space, which enables us to handle Gaussians with Euclidean operations instead of complicated Riemannian operations while respecting the geometry of Gaussians. The proposed descriptor, called Local Log-Euclidean Multivariate Gaussian (L^{2}EMG) descriptor, can model statistics of both low-dimensional raw features and high-dimensional ones. Unlike the popular histogram-based descriptors such as SIFT and HoG which estimate zero-order statistics by quantizing the feature spaces, L^{2}EMG is continuous and can model higher-order statistics.

Fig. 1(a) illustrates the flowchart of our L^{2}EMG descriptor. Given an input image, we first extract n-dimensional raw features, then compute multivariate Gaussians at a dense grid, obtaining a function-valued image, and finally, embed these Gaussians into a linear space to obtain the vectorized L^{2}EMG descriptors. We develop two embedding methods, as shown in Figs. 1(b) and 1(c), respectively. The first method, called direct embedding Log-Euclidean (DE-LogE), maps $A^+(n+1)$ via matrix logarithm to the linear space $A(n+1)$. The second one, what we call indirect embedding Log-Euclidean (IE-LogE), first maps $A^+(n+1)$ via the coset and polar decomposition into the space of symmetric positive definite (SPD) matrices, $Sym^+(n+1)$, and then into the linear space $Sym(n+1)$ by the Log-Euclidean framework [14]. DE-LogE has a simple, analytic expression for diagonal-covariance Gaussians, particularly suitable for high-dimensional raw features. Our methods are primarily based on Lie group isomorphisms, which respects the algebraic and topological structures of the spaces involved.

This paper substantially extends our previous work namely the local Log-Euclidean covariance matrix (L^{2}ECM) [15] from three aspects. (1) We represent the local image statistics by using multivariate Gaussians instead of using only covariance matrices as in L^{2}ECM. L^{2}ECM is a special case of L^{2}EMG where the mean vectors are always constant. Hence, L^{2}EMG can capture richer statistical information. (2) We equip the space of Gaussians with a Lie group structure and present novel methods to deal with Gaussians by using the Euclidean operations rather than the Riemannian operations. (3) We evaluate thoroughly the L^{2}EMG descriptor and compare it with L^{2}ECM as well as other popular descriptors on a variety of benchmark datasets, including object classification, scene categorization and material recognition.

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 introduces some basics of Lie groups. Section 4 presents the Lie group structure of the space of Gaussians and the methods to embed Gaussians in the linear spaces. Section 5 describes the computation of the proposed descriptors and the complexity analysis. Section 6 presents experiments to evaluate and analyze our descriptors. Finally, Section 7 concludes the paper and discusses future work.

2 Related Work

This section begins with a review of image descriptors and then introduces the existing methods to handle Gaussians by means of information geometry.

2.1 Image Descriptors

We focus on distribution-based descriptors which have achieved state-of-the-art performance in image classification (cf. [1] for other types of descriptors such as those based on spatial-frequency analysis or image moments).

2.1.1 Histogram-based descriptors

The SIFT descriptor which builds three-dimensional histograms of spatial cells and gradient orientation [9] is one of the representatives in this category. PCA-SIFT [16] packs the gradients in an image patch by using principal component analysis (PCA) for dimensionality reduction. Extensions of SIFT by considering color invariance improve the discriminative capability but increase the size of descriptors [17]. In [1], image patches are divided into log-polar cells rather than Cartesian ones to extract the Gradient Location and Orientation Histogram (GLOH). Histogram of Orientation Gradient (HOG) [10] is originally proposed for human detection and now widely used in many vision tasks. SURF [18] and DAISY [19] descriptors enjoy computational efficiency while preserving the strengths of SIFT and GLOH. Chen et al. [20], inspired by the law of the perception of human beings, proposed the Weber local descriptor (WLD) by building 2D histograms of differential excitation and gradient orientation.

Another line of research employs the order measure to improve robustness to complex, nonlinear intensity variations induced by illumination or lighting changes. The local binary pattern (LBP) [21], based on the order relation of image intensity in a local neighborhood, is intensity and rotation invariant. The CENTRIST estimates histogram of Census Transform values and it works well for place and scene recognition [22]. The locally stable monotonic change invariant feature descriptor (SMD) [23] models intensity orders of pairs of pixels. The ordinal spatial intensity distribution (OSID) descriptor [24] is a 2D histogram of intensity orderings and spatial sub-division spaces.

2.1.2 Probability density function based descriptors

Tuzel et. al [7] proposed the covariance descriptor and adopted the Affine-invariant Riemannian metric [25] since the space of covariance matrices is not a vector space but forms a Riemannian manifold. The L^{2}ECM descriptors [15] consist of pixel-wise covariance matrices to represent the local feature correlations, which are embedded in the linear space through the Log-Euclidean metric [14]. In object tracking, Gong et al. [26] proposed the shape of Gaussian descriptor utilizing both the mean vector and covariance matrix. In scene categorization, Nakayama et al. [27] modeled a whole image with a Gaussian distribution and measured the dissimilarity based on the α-divergence [28]. In [29], the local high-order statistics are employed by using the Fisher Vector method [4]. Giuseppe et al. [30] adopted multivariate Gaussians as image descriptors, in which the mean vector is concatenated with the covariance matrix mapped to the tangent space through the Affine-invariant Riemannian metric. Ma et al. [31] described the human body regions with multivariate Gaussians and evaluated the distance based on the product of Lie groups [32].

2.2 Gaussian Embedding

The space $N(n)$ of Gaussians can be naturally seen as a Riemannian manifold equipped with the Fisher metric [33]. However, the geodesics distance on $N(n)$ has no closed form, except for univariate Gaussians or multivariate ones with fixed mean vectors or fixed covariance matrices [34]. Amari et al. [28] established a dually-flat structure to handle the manifolds of probability distributions, i.e., two flat affine coordinate systems which are mutually orthogonal with respect to the Fisher-Rao metric. Therein α-divergence is proposed to evaluate the dissimilarity between distributions and it is equivalent to the Kullback-Leibler (KL) divergence when $\alpha = -1$. In [26], the space of Gaussians is embedded into an affine group and the Riemannian metric is adopted to measure the distance. Calvo et al. [35] and Lovrić et al. [36] identified Gaussians as SPD matrices by embedding Gaussians in the Siegel group or the Riemannian symmetric space, respectively. They also studied the Riemannian metric between SPD matrices.

Table 2 summarizes and compares various embedding methods. We have the following observations. (1) Almost all the existing methods regard $N(n)$ as a Riemannian manifold and fail to explore its algebraic structure. Note that [26], [35], [36] showed that the space of Gaussians can be embedded into some groups. In contrast, we show that the space of Gaussians *itself* can be considered as a Lie group by defining a multiplication operation directly on the manifold formed by Gaussians (Section 4.1), i.e., a group with smooth group multiplication and inverse operations. This gives us insights into the algebraic and geometrical structure of the space of Gaussians. (2) When evaluating the distances [26], [35], [36] or dissimilarity measures [27], [28], the involved matrices in these methods entangle, making them hard to handle Gaussians conveniently and efficiently, particularly for large-scale problems. Based on the Lie group structure of $N(n)$, we further propose novel methods to embed $N(n)$ in linear spaces. Our embedding processes depend primarily on Lie group isomorphisms, establishing equivalences between the corresponding spaces. Consequently, we can handle Gaussians with Euclidean operations instead of Riemannian operations, while respecting the geometry of Gaussians. Compared to previous works, our metric is untangled and can handle efficiently large-scale problem.

3 A Brief Introduction on Lie Group

This section introduces some basic background of Lie group and matrix group; for complete theory one may refer to textbooks such as [37], [38].

TABLE 2
Comparison of embedding methods for Gaussian distributions. Here $\widetilde{\mathbf{L}}$ and $\mathbf{L}$ denote the Cholesky factors of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$, respectively, while Gaussians $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ are respectively identified as $\mathbf{B}_1$ and $\mathbf{B}_2$.

	Space of Gaussians	Embedding method	Embedding form $\mathbf{B}$ of $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$	Metric or (dis)similarity	Decoupling
Skovgaard et al. [34]	Riemannian manifold	Embeded in an open subset of the Euclidean space	$[(\mu_i)_{i=1,\ldots,n}, (\sigma_{ij})_{i\le j}]$	Fisher information metric (No closed-form in general)	×
Amari et al. [27], [28]	Riemannian manifold	Embeded in a flat manifold by taking an affine coordinate system	$[(\mu_i)_{i=1,\ldots,n}, (\sigma_{ij} + \mu_i\mu_j)_{i\le j}]$	α-divergence (α=1) $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ $+\mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2-\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1})-2n$	×
Gong et al. [26]	Riemannian manifold	Embeded in an affine group	$\mathbf{B} = \begin{bmatrix} \widetilde{\mathbf{L}} & \boldsymbol{\mu} \\ \mathbf{0}^T & 1 \end{bmatrix}$	$\lVert \log(\mathbf{B}_1^{-1}\mathbf{B}_2) \rVert_F$	×
Calvo et al. [35]	Riemannian manifold	Embeded via a diffemorphism in a Siegel group	$\mathbf{B} = \begin{bmatrix} \mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix}$	$\lVert \log(\mathbf{B}_1^{-1}\mathbf{B}_2) \rVert_F$	×
Lovrić et al. [36]	Riemannian manifold	Embeded via group action in a Riemannian symmetric s-pace	$\mathbf{B} = \lvert \mathbf{\Sigma} \rvert^{-\frac{2}{n+1}} \begin{bmatrix} \mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix}$	$\lVert \log(\mathbf{B}_1^{-1}\mathbf{B}_2) \rVert_F$	×
L^{2}EMG	Lie group	Embeded in the linear spaces via Lie group theory	$\mathbf{B} = \log \begin{bmatrix} \mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix}^{\frac{1}{2}}, \log \begin{bmatrix} \mathbf{L}^{-T} & \boldsymbol{\mu} \\ \mathbf{0}^T & 1 \end{bmatrix}$	$\lVert \mathbf{B}_1-\mathbf{B}_2 \rVert_F$	√

3.1 Lie Group

A *group* G is a set equipped with a multiplication operation $\cdot: G \times G \mapsto G$, which combines any two elements a and b in G to produce an element, written $a \cdot b$, of G, and the following properties are satisfied:

(1) The multiplication operation is associative. For all a, b and c in G, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.
(2) There is an identity element e such that for each a in G $a \cdot e = e \cdot a = a$.
(3) For each element a in G, there is an inverse a^{-1} for which $a \cdot a^{-1} = a^{-1} \cdot a = e$.

For all a, b in G, if $a \cdot b = b \cdot a$, then G is said to be a commutative (or abelian) group. A subset H of G is called a *subgroup* of G if H forms a group under the operation $\cdot$. A *Lie group* is a group that is also a differential manifold, with the property that the group multiplication and inverse are smooth functions. A Lie group is locally equivalent to a linear space and thus the local neighborhood of any element can be adequately described by its tangent space. The tangent space of the identity element forms a *Lie algebra*, a linear space with a bilinear product called Lie bracket.

Let G be a Lie group. A *Lie subgroup* H of G is a closed subgroup of G which is also a submanifold. A *left coset* of H in G is a subset of the form

$$aH = \{a \cdot h | h \in H\},$$

where $a \in G$. aH and bH are equal if they have an element in common. The set of all the cosets of H, denoted by G/H, forms a partition of the group G, i.e., G is the union of all distinct cosets of H. One can also define the right coset of H in G, i.e., $Ha = \{h \cdot a | h \in H\}$, where $a \in G$ and the aforementioned arguments hold similarly.

Let G, G' be Lie groups and $\cdot, \circ$ be their corresponding multiplication operations. A *Lie group homomorphism* $\phi : G \to G'$ is a smooth function that satisfies

$$\phi(a \cdot b) = \phi(a) \circ \phi(b) \quad \text{for all } a, b \in G.$$

If, in addition, ϕ is a bijective function (one to one and onto) and the inverse mapping ϕ^{-1} is smooth, then we say that ϕ is a *Lie group isomorphism* or G is *isomorphic* to G'. For isomorphic Lie groups, the operation in one Lie group is smoothly carried over to the operation in another. Since having the same algebraic and topological properties, isomorphic Lie groups are equivalent.

3.2 Matrix Group

The most interesting examples of Lie groups in computer vision are those formed by square matrices. The set of all $n \times n$ real (resp. complex) invertible matrices under matrix multiplication forms a Lie group called real (resp. complex) general linear group. The matrix groups are Lie subgroups of the general linear groups. The Lie algebra of a matrix group G, denoted by $\mathfrak{g}$, is the set of all matrices $\mathbf{X}$ such that the one-parameter group $\exp(z\mathbf{X})$ is in G for any real number z [37, Section 2.5]. In this paper we restrict ourselves to square matrices over the field $\mathbb{R}$ of real numbers, unless otherwise stated, since they are of our most interest. The following notations will be used: $GL^+(n)-$the group of $n \times n$ matrices with positive determinant; $SO(n)-$the special orthogonal group, i.e., the group of $n \times n$ orthogonal matrices of determinant one; $PDUT(n)-$the group of $n \times n$ upper triangular matrices with positive diagonal entries; $Sym^+(n)-$the group of $n \times n$ SPD matrices. The set of all $n \times n$ upper triangular matrices, denoted by $Ut(n)$, and the set of all $n \times n$ symmetric matrices, denoted by $Sym(n)$, are Lie algebras of $PDUT(n)$ and $Sym^+(n)$, respectively.

The matrix exponential and logarithm play a crucial role in the matrix group theory. Let $\mathbf{X}$ be a square matrix. The matrix exponential of $\mathbf{X}$, denoted by $\exp(\mathbf{X})$, generalizes the scalar exponential and is defined by the power series

$$\exp(\mathbf{X}) = \sum_{k=0}^{\infty} \frac{\mathbf{X}^k}{k!}. \tag{1}$$

The series converge for any $\mathbf{X}$ and $\exp(\mathbf{X})$ is smooth [14]. The logarithm of a matrix $\mathbf{A}$, denoted by $\log(\mathbf{A})$, is a matrix $\mathbf{X}$ such that $\exp(\mathbf{X}) = \mathbf{A}$. In the field $\mathbb{C}$ of complex numbers, the logarithm of any invertible matrix exists but may not be unique [37, Theorem 2.9]. For a real invertible matrix $\mathbf{A}$, there exists a real logarithm if and only if the number of elementary divisors belonging to each negative eigenvalue is even, while $\mathbf{A}$ has a unique real logarithm if it has no negative eigenvalues [39, Theorems 3.4 and 3.11].

3.3 Log-Euclidean on $GL^+(1)$

It is well known that $GL^+(1)$ is a Lie group, and it is indeed equivalent to the Lie group $\mathbb{R}^+$ formed by all positive real numbers under multiplication. The Lie algebra of $\mathbb{R}^+$ is the set $\mathbb{R}$ of all real numbers. Here we formulate this fact in the Log-Euclidean notation[1] by means of the following two Propositions. We underline that this notation is fundamental and can be extended to a more general scenario.

Proposition 1. *The exponential of real numbers*

$$\exp : \mathbb{R} \to \mathbb{R}^+, x \mapsto \exp(x)$$

is a smooth bijection and its inverse log *is also smooth.*

Proposition 2 (Log-Euclidean on $\mathbb{R}^+$)**.** *We define*

$$\otimes : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+,\, a_1 \otimes a_2 = \exp(\log(a_1) + \log(a_2))$$
$$\text{and } \odot : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+,\, \lambda \odot a = \exp(\lambda \log(a)) = a^\lambda. \quad (2)$$

Under operation $\otimes$, $\mathbb{R}^+$ is a commutative Lie group, and $\log : \mathbb{R}^+ \to \mathbb{R}$ *is a Lie group isomorphism, i.e.,* $\log(a_1 \otimes a_2) = \log(a_1) + \log(a_2)$. *$\mathbb{R}^+$ is a linear space under $\otimes$ and $\odot$.*

Proposition 1 is well-known and the exponential function exp and its inverse log are both diffeomorphisms. The proof of Proposition 2 can be readily completed by the standard definition of Lie group.

Proposition 2 states that through the logarithm, the multiplications in the Lie group $\mathbb{R}^+$ are transformed to the additions in the logarithm domain (linear space $\mathbb{R}$). In this sense, we call this methodology "Log-Euclidean". The Log-Euclidean method on $Sym^+(n)$ proposed by Arsigny et al. [14] can be viewed as an extension of Proposition 2 from $\mathbb{R}^+$ to $Sym^+(n)$. In Section 4.2.1 we will extend this proposition to handle the space of Gaussians.

4 Structure of Space of Gaussians and Embedding

In this section, we first show that the space of Gaussians can be provided with a Lie group structure and then describe two methods to embed Gaussians in linear spaces.

4.1 Lie Group Structure of Gaussians

Let $\mathbf{\Sigma}$ be any $n \times n$ SPD matrix and $\mathbf{\Sigma}^{-1}$ be its inverse. We know that $\mathbf{\Sigma}^{-1}$ is also an SPD matrix and it has a unique Cholesky decomposition $\mathbf{\Sigma}^{-1} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is upper triangular with positive diagonal entries. Hence, it follows that $\mathbf{\Sigma} = \mathbf{L}^{-T}\mathbf{L}^{-1}$, where $\mathbf{L}^{-T} \in PDUT(n)$ denotes the transpose of the inverse of $\mathbf{L}$. That is, any SPD matrix can be uniquely decomposed as the product of an upper triangular matrix with positive diagonal entries and its transpose.

Based on the decomposition above, we define a multiplication operation on $N(n)$.

Definition 1. Let $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \in N(n), i = 1, 2$, be two arbitrary Gaussians and $\mathbf{\Sigma}_i = \mathbf{L}_i^{-T}\mathbf{L}_i^{-1}$, where $\mathbf{L}_i$ is the Cholesky factor of $\mathbf{\Sigma}_i^{-1}$. We define an operation $\star$ between two Gaussians as

$$\begin{aligned} \star :& N(n) \times N(n) \to N(n), \qquad (3)\\ &\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1) \star \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)\\ &\quad = \mathcal{N}(\mathbf{L}_1^{-T}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1, (\mathbf{L}_1\mathbf{L}_2)^{-T}(\mathbf{L}_1\mathbf{L}_2)^{-1}). \end{aligned}$$

We may interpret (3) in the following manner. Suppose that random vector $\mathbf{x}$ has the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$, i.e., $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$. The operation $\star$ can be viewed as an affine transformation $(\mathbf{L}_1^{-T}, \boldsymbol{\mu}_1)$ acting on $\mathbf{x}$ and the resulting random vector $\mathbf{y} = \mathbf{L}_1^{-T}\mathbf{x} + \boldsymbol{\mu}_1$ follows Gaussian distribution $\mathcal{N}(\mathbf{L}_1^{-T}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1, (\mathbf{L}_1\mathbf{L}_2)^{-T}(\mathbf{L}_1\mathbf{L}_2)^{-1})$.

Upon Definition 1, we have the following theorem:

Theorem 1. $N(n)$ is a Lie group under multiplication operation $\star$ as defined in (3).

Proof Let $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \in N(n), i = 1, 2, 3$, be three arbitrary Gaussians and $\mathbf{\Sigma}_i = \mathbf{L}_i^{-T}\mathbf{L}_i^{-1}$, where $\mathbf{L}_i$ is the Cholesky factor of $\mathbf{\Sigma}_i^{-1}$. The operation $\star$ is associative since the multiplication of the three Gaussians with either groupings are equal, i.e.,

$$\begin{aligned} &(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1) \star \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)) \star \mathcal{N}(\boldsymbol{\mu}_3, \mathbf{\Sigma}_3)\\ &\quad = \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1) \star (\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2) \star \mathcal{N}(\boldsymbol{\mu}_3, \mathbf{\Sigma}_3))\\ &\quad = \mathcal{N}((\mathbf{L}_1\mathbf{L}_2)^{-T}\boldsymbol{\mu}_3 + \mathbf{L}_1^{-T}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1, (\mathbf{L}_1\mathbf{L}_2\mathbf{L}_3)^{-T}(\mathbf{L}_1\mathbf{L}_2\mathbf{L}_3)^{-1}). \end{aligned}$$

The standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{0}$ and $\mathbf{I}$ denote the zero vector and identity matrix, respectively, is the unique identity element in $N(n)$, i.e.,

$$\mathcal{N}(\mathbf{0}, \mathbf{I}) \star \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \star \mathcal{N}(\mathbf{0}, \mathbf{I}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}).$$

Any Gaussian $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \in N(n)$ has an inverse given by

$$\mathcal{N}^{-1}(\boldsymbol{\mu}, \mathbf{\Sigma}) = \mathcal{N}(-\mathbf{L}^T\boldsymbol{\mu}, \mathbf{L}^T\mathbf{L}) \qquad (4)$$

where $\mathbf{\Sigma} = \mathbf{L}^{-T}\mathbf{L}^{-1}$ and $\mathbf{L}^{-T} \in PDUT(n)$. In light of the uniqueness of both the Cholesky decomposition and the inversion of SPD matrix, we know that the inverse function $\mathcal{N}^{-1}$ is unique. Hence, $N(n)$ is a group.

According to the Cholesky decomposition algorithm [40, Section 4.2], each entry of $\mathbf{L}$ can be written as the composite of arithmetic operations of addition (subtraction), multiplication (division) or the square root on the entries of $\mathbf{\Sigma}^{-1}$. We thus conclude that the Cholesky decomposition is smooth [38, Proposition 1.13]. The smooth properties of matrix multiplication and matrix inversion can be found in [38, Section 1.2]. Hence, the multiplication operation (3) and inverse mapping (4) defined on the manifold of Gaussians are both smooth. Therefore, $N(n)$ is a Lie group. □

Let

$$A^+(n+1) = \left\{ \mathbf{A}_{\boldsymbol{\mu},\mathbf{Z}} \triangleq \begin{bmatrix} \mathbf{Z} & \boldsymbol{\mu} \\ \mathbf{0}^T & 1 \end{bmatrix} | \mathbf{Z} \in PDUT(n), \boldsymbol{\mu} \in \mathbb{R}^n \right\}.$$

Obviously $A^+(n+1)$ is a closed subgroup of $GL^+(n+1)$, and so it is a Lie group. The following theorem establishes the equivalence between $N(n)$ and $A^+(n+1)$.

Theorem 2. The function

$$\phi :\ A^+(n+1) \to N(n),\ \ \phi(\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}), \qquad (5)$$

where $\mathbf{\Sigma} = \mathbf{L}^{-T}\mathbf{L}^{-1}$ and $\mathbf{L}^{-T} \in PDUT(n)$, is a Lie group isomorphism.

Proof Obviously ϕ is a smooth function as the matrix multiplication and exponential function are smooth, and the smoothness of its inverse follows from the fact that the decomposition $\mathbf{\Sigma} = \mathbf{L}^{-T}\mathbf{L}^{-1}$ is smooth. Through ϕ, $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$ is uniquely mapped to Gaussian $\mathcal{N}(\boldsymbol{\mu}, \mathbf{L}^{-T}\mathbf{L}^{-1})$. For any $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) \in N(n)$, as $\mathbf{\Sigma}$ has a unique decomposition $\mathbf{\Sigma} = \mathbf{L}^{-T}\mathbf{L}^{-1}, \mathbf{L}^{-T} \in PDUT(n)$, the inverse

[1] Note that the notation "Log-Euclidean" was first used in [14] and our formulation is highly inspired by this work.

function ϕ^{-1} uniquely exists. Consider two arbitrary Guassians $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in N(n), i = 1, 2$ and their decompositions $\boldsymbol{\Sigma}_i = \mathbf{L}_i^{-T}\mathbf{L}_i^{-1}$, $\mathbf{L}_i^{-T} \in PDUT(n)$. It is straightforward to show that ϕ is compatible with the law of composition, i.e., $\phi(\mathbf{A}_{\boldsymbol{\mu}_1,\mathbf{L}_1^{-T}}\mathbf{A}_{\boldsymbol{\mu}_2,\mathbf{L}_2^{-T}}) = \phi(\mathbf{A}_{\boldsymbol{\mu}_1,\mathbf{L}_1^{-T}}) \star \phi(\mathbf{A}_{\boldsymbol{\mu}_2,\mathbf{L}_2^{-T}})$. Hence, ϕ is a Lie group isomorphism. □

Through ϕ^{-1}, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is identified as an upper triangular matrix $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$, whose diagonal entries are positive with the last being one. In contrast, the embedding matrix in [26] does not have such a desirable form.

4.2 Embedding Gaussians in Linear Space

4.2.1 Direct Embedding (DE-LogE)

Let us consider the set

$$A(n+1) = \left\{ \mathbf{A}_{\mathbf{t},\mathbf{X}} \triangleq \begin{bmatrix} \mathbf{X} & \mathbf{t} \\ \mathbf{0}^T & 0 \end{bmatrix} | \mathbf{X} \in Ut(n), \mathbf{t} \in \mathbb{R}^n \right\}. \quad (6)$$

In terms of the definitions of matrix exponential (1) and its property [37, Proposition 2.4], as well as the properties of triangular matrix, one can show that $A(n+1)$ is the Lie algebra of the matrix group $A^+(n+1)$.

The following theorem states that $\exp$ is a diffeomorphism from $A^+(n+1)$ to its Lie algebra $A(n+1)$.

Theorem 3. The function

$$\exp : A(n+1) \to A^+(n+1), \mathbf{A}_{\mathbf{t},\mathbf{X}} \mapsto \exp(\mathbf{A}_{\mathbf{t},\mathbf{X}})$$

is a smooth bijection and its inverse is smooth as well.

Proof It is not difficult to know that any $\mathbf{A}_{\mathbf{t},\mathbf{X}}$ in $A(n+1)$ is uniquely mapped through the exponential function to $A^+(n+1)$; conversely, any $\mathbf{A}_{\boldsymbol{\mu},\mathbf{Z}} \in A^+(n+1)$ has positive eigenvalues and thus $\log(\mathbf{A}_{\boldsymbol{\mu},\mathbf{Z}})$ uniquely exists in $A(n+1)$ [39], [41]. So $\exp : A(n+1) \to A^+(n+1)$ is one to one and onto, while the smoothness of $\exp$ and that of its inverse are guaranteed by [39, Theorem 3.11]. □

Now we can extend the Log-Euclidean framework to $A^+(n+1)$ by the following theorem.

Theorem 4 (Log-Euclidean on $A^+(n+1)$)**.** We define

$$\begin{aligned} \otimes : & A^+(n+1) \times A^+(n+1) \to A^+(n+1), \\ & \mathbf{A}_1 \otimes \mathbf{A}_2 = \exp(\log(\mathbf{A}_1) + \log(\mathbf{A}_2)), \text{ and} \\ \odot : & \mathbb{R} \times\ A^+(n+1) \to A^+(n+1), \\ & \lambda \odot \mathbf{A} = \exp(\lambda \log(\mathbf{A})) = \mathbf{A}^\lambda. \end{aligned} \quad (7)$$

Under operation $\otimes$, $A^+(n+1)$ is a commutative Lie group,

$$\log : A^+(n+1) \to A(n+1), \mathbf{A} \mapsto \log(\mathbf{A}) \quad (8)$$

is a Lie group isomorphism. In addition, under $\otimes$ and $\odot$, $A^+(n+1)$ is a linear space.

Proof of this theorem is straightforward and is therefore omitted [note that $A(n+1)$ is a Lie group under *matrix addition*]. By far $A^+(n+1)$ is equipped with a novel Lie group structure. The isomorphism establishes the equivalence between $A^+(n+1)$ and $A(n+1)$, which indicates that the operations on $A^+(n+1)$ can be transformed, via matrix logarithm, to the linear space $A(n+1)$ while respecting the algebraic and topological structure of $A^+(n+1)$. It is worth mentioning that, according to [39, Theorem 3.11], we can draw a more general conclusion as follows. Let $IN(n)$ be the set of all $n \times n$ real invertible matrices with nonnegative eigenvalues. Since any $\mathbf{S} \in IN(n)$ has a unique real logarithm, and $\exp : \log(IN(n)) \to IN(n)$ is a diffeomorphsim, where $\log(IN(n))$ denotes the image of $IN(n)$ under logarithm, we can establish the Log-Euclidean on $IN(n)$.

We finally illustrate the complete embedding process of DE-LogE as follows:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \overset{\phi^{-1}}{\Longrightarrow} \mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}} \overset{\log}{\Longrightarrow} \log(\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}), \quad (9)$$

where $\boldsymbol{\Sigma} = \mathbf{L}^{-T}\mathbf{L}^{-1}$ and $\mathbf{L}^{-T} \in PDUT(n)$. Recall that $\mathbf{L}$ is the Cholesky factor of $\boldsymbol{\Sigma}^{-1}$.

4.2.2 Indirect Embedding (IE-LogE)

This embedding method consists of three consecutive functions. Let us consider the left coset of $SO(n+1)$ in $GL^+(n+1)$

$${}_{\mathbf{P}}SO = \{\mathbf{PO} | \mathbf{O} \in SO(n+1)\}, \quad (10)$$

where $\mathbf{P} \in Sym^+(n+1)$ is an $(n+1) \times (n+1)$ SPD matrix. As $|\mathbf{PO}| = |\mathbf{P}||\mathbf{O}| = |\mathbf{P}| > 0$, where $|\cdot|$ denotes the matrix determinant, ${}_{\mathbf{P}}SO$ is a subset of $GL^+(n+1)$. Conversely, any matrix $\mathbf{G} \in GL^+(n+1)$ has a unique left polar decomposition [42] $\mathbf{G} = \mathbf{PR}$, where $\mathbf{P} \in Sym^+(n+1)$ and $\mathbf{R} \in SO(n+1)$, and thus $\mathbf{G}$ belongs to one and only one coset ${}_{\mathbf{P}}SO$. Hence, the set of cosets $\{{}_{\mathbf{P}}SO, \mathbf{P} \in Sym^+(n+1)\}$ partitions $GL^+(n+1)$ and we denote the set by the quotient $GL^+(n+1)/SO(n+1)$ which is well-known to be a Lie group [14]. Note that $A^+(n+1)$ is a Lie subgroup of $GL^+(n+1)$, and there exists an injective function for which

$$\begin{aligned} \pi : \ & A^+(n+1) \to GL^+(n+1)/SO(n+1), \\ & \pi(\mathbf{A}) = {}_{\mathbf{P}}SO, \end{aligned} \quad (11)$$

where $\mathbf{A} = \mathbf{PR}$ is the left polar decomposition of $\mathbf{A}$.

Next, we map, through the bijective function

$$\begin{aligned} \gamma : \ & GL^+(n+1)/SO(n+1) \to Sym^+(n+1), \\ & \gamma({}_{\mathbf{P}}SO) = \mathbf{P} \end{aligned} \quad (12)$$

the coset ${}_{\mathbf{P}}SO$ to the space of SPD matrices $Sym^+(n+1)$. Below we show that γ is a Lie group isomorphism by defining an operation

$$\begin{aligned} * : \ & GL^+(n+1)/SO(n+1) \times GL^+(n+1)/SO(n+1) \\ & \to GL^+(n+1)/SO(n+1) \\ & {}_{\mathbf{P}}SO *_{\mathbf{Q}} SO = {}_{\mathbf{P}\odot\mathbf{Q}}SO, \end{aligned} \quad (13)$$

where $\mathbf{P} \odot \mathbf{Q} = \exp(\log(\mathbf{P}) + \log(\mathbf{Q}))$ is the logarithmic multiplication defined on $Sym^+(n+1)$ [14]. The function γ is a Lie group homomorphism since $\gamma({}_{\mathbf{P}}SO *_{\mathbf{Q}} SO) = \gamma({}_{\mathbf{P}\odot\mathbf{Q}}SO) = \mathbf{P} \odot \mathbf{Q} = \gamma({}_{\mathbf{P}}SO) \odot \gamma({}_{\mathbf{Q}}SO)$. The smoothness of γ and that of its inverse are obvious and thus it is a Lie group isomorphism, which guarantees that $GL^+(n+1)/SO(n+1)$ be equivalent to $Sym^+(n+1)$.

The third function aims to map the SPD matrices into a linear space. To this end, we adopt the Log-Euclidean framework proposed by Arsigny et al. [14]. Their idea is to transform, via matrix logarithm, the Riemannian operations on $Sym^+(n+1)$ to the Euclidean ones in the vector space $Sym(n+1)$; we refer the readers to [14] for details.

Theorem 5 (Log-Euclidean on $Sym^+(n+1)$)**.** Under operation $\otimes$, $Sym^+(n+1)$ is a commutative Lie group, and

$$\log:\ Sym^+(n+1)\to Sym(n+1),\ \mathbf{P}\mapsto\log(\mathbf{P}) \qquad (14)$$

is a Lie group isomorphism. In addition, under $\otimes$ and $\odot$, $Sym^+(n+1)$ is a linear space.

Thus far, we summarize the complete embedding process of IE-LogE as follows:

$$\begin{tikzcd}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \arrow[r, Rightarrow, "\phi^{-1}"] & \mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}} \arrow[r, Rightarrow, "\pi"] & {}_{\mathbf{P}}SO(n+1) \arrow[d, Rightarrow, "\gamma"] \\ & \log(\mathbf{P}) & \mathbf{P} \arrow[l, Rightarrow, "\log"']\end{tikzcd} \qquad (15)$$

Here $\boldsymbol{\Sigma}=\mathbf{L}^{-T}\mathbf{L}^{-1}$ and $\mathbf{L}$ is the Cholesky factor of $\boldsymbol{\Sigma}^{-1}$; $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$ has left polar decomposition $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}=\mathbf{PR}$.

Properties of P in (15) This matrix has two properties.

1) It is the square root matrix of $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}\mathbf{A}^T_{\boldsymbol{\mu},\mathbf{L}^{-T}}$, i.e.,

$$\mathbf{P}=\begin{bmatrix}\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu}\\ \boldsymbol{\mu}^T & 1\end{bmatrix}^{\frac{1}{2}}. \qquad (16)$$

The eigenvalues of $\mathbf{P}$ are identical to the singular values of $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$. In addition, their ℓ_2-norm condition numbers are identical.

2) The matrix $\mathbf{R}$ that accompanies $\mathbf{P}$ is the closest possible orthogonal matrix to $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$ [43]. That is

$$\mathbf{R}=\arg\min\nolimits_{\mathbf{O}\in O(n+1)}\|\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}-\mathbf{O}\|_F, \qquad (17)$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $O(n+1)$ is the orthogonal group of dimension $n+1$.

Embedding by right coset In previous development we have accomplished the embedding of Gaussians based on the left coset of $SO(n+1)$. In a very similar manner, we can consider the right coset $SO_{\mathbf{P}}=\{\mathbf{OP}|\mathbf{O}\in SO(n+1)\}$ to obtain the second embedding scheme. The space of cosets $\{SO_{\mathbf{P}},\mathbf{P}\in Sym^+(n+1)\}$ partitions $GL^+(n+1)$, and we denote this space by $GL^+(n+1)\backslash SO(n+1)$. Note that any invertible matrix has a unique *right* polar decomposition [43]. We map a matrix $\mathbf{A}\in A^+(n+1)$ to an SPD matrix through the following two functions:

$$\tilde{\pi}:A^+(n+1)\to GL^+(n+1)\backslash SO(n+1),\tilde{\pi}(\mathbf{A})=SO_{\mathbf{P}'},$$
$$\tilde{\gamma}:GL^+(n+1)\backslash SO(n+1)\to Sym^+(n+1),\tilde{\gamma}(SO_{\mathbf{P}'})=\mathbf{P}'.$$

Here $\mathbf{A}=\mathbf{R}'\mathbf{P}',\mathbf{R}'\in SO(n+1),\mathbf{P}'\in Sym^+(n+1)$ is the right polar decomposition of $\mathbf{A}$. In this case, the embedding matrix $\mathbf{P}'$ is the square root of $\mathbf{A}^T_{\boldsymbol{\mu},\mathbf{L}^{-T}}\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$, i.e.,

$$\mathbf{P}'=\begin{bmatrix}\mathbf{L}^{-1}\mathbf{L}^{-T} & \mathbf{L}^{-1}\boldsymbol{\mu}\\ \boldsymbol{\mu}^T\mathbf{L}^{-T} & \boldsymbol{\mu}^T\boldsymbol{\mu}+1\end{bmatrix}^{\frac{1}{2}}. \qquad (18)$$

Based on the left coset and right coset, we obtain the SPD matrices (16) and (18), respectively. Interestingly, Eq. (16) shares a similar form to those in [35], [36]. However, our embedding mechanisms are different from theirs; most importantly, we further map SPD matrices into the linear space to handle Gaussians with Euclidean operations.

5 Computation of L^{2}EMG Descriptor

In this section, we describe the process of computing L^{2}EMG descriptors and analyze its complexity.

5.1 Estimation of Local Gaussians

Let I be an input image and $\mathbf{f}(\mathbf{z})$ be the n-dimensional vector of raw features computed at the spatial coordinate $\mathbf{z}=(x,y)$. We compute a function-valued image in which each pixel $\mathbf{z}$ is represented by a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}(\mathbf{z}),\boldsymbol{\Sigma}(\mathbf{z}))$. The Gaussian can be estimated via the maximum likelihood method in a local image region. Let $G_r(\mathbf{z})$ be a $r\times r$ image patch centered at $\mathbf{z}$. The estimated Gaussian can be written as

$$\begin{aligned}\mathcal{N}(\mathbf{z})&\triangleq\mathcal{N}(\boldsymbol{\mu}(\mathbf{z}),\boldsymbol{\Sigma}(\mathbf{z})) \qquad (19)\\ &=\left|2\pi\boldsymbol{\Sigma}(\mathbf{z})\right|^{-\frac{1}{2}}\exp(-(\mathbf{f}-\boldsymbol{\mu}(\mathbf{z}))^T\boldsymbol{\Sigma}(\mathbf{z})^{-1}(\mathbf{f}-\boldsymbol{\mu}(\mathbf{z}))),\end{aligned}$$

where $|\cdot|$ denotes the matrix determinant, and

$$\begin{aligned}\boldsymbol{\mu}(\mathbf{z})&=\frac{1}{r^2}\sum_{\mathbf{z}'\in G_r(\mathbf{z})}\mathbf{f}(\mathbf{z}') \qquad (20)\\ \boldsymbol{\Sigma}(\mathbf{z})&=\frac{1}{r^2-1}\sum_{\mathbf{z}'\in G_r(\mathbf{z})}(\mathbf{f}(\mathbf{z}')-\boldsymbol{\mu}(\mathbf{z}))(\mathbf{f}(\mathbf{z}')-\boldsymbol{\mu}(\mathbf{z}))^T\end{aligned}$$

are the empirical mean vector and sample covariance matrix, respectively. In practice, some $\boldsymbol{\Sigma}(\mathbf{z})$ may be rank-deficient and so we add a small positive number to the diagonal entries of each covariance matrix. A small r is helpful in capturing fine scale local structure, while with a big r the local statistics at larger scales will be captured.

The function-valued image can be obtained by constructing the integral images for mean vectors and covariance matrices [7]. Given a raw feature map, we build respectively an integral image for each component of the raw feature vector, and an integral image for the cross product of any pair of components. As each integral image can be computed via one pass of the original one, the computational cost to build all integral images is $O(n(n+3)|I|)$, where $|I|$ denotes the area of the image. Through these integral images, the empirical mean vector and covariance matrix of a rectangular patch of any size can be obtained using $2n(n+3)$ additions and two divisions.

5.2 Computation of Embedding Matrix

In what follows, we describe how to compute the embedding matrices. We omit the spatial coordinate $\mathbf{z}$ in the mean vector and covariance matrix for the convenience of expression. To facilitate operations, we vectorize the final embedding matrices.

DE-LogE For the decomposition $\boldsymbol{\Sigma}=\mathbf{L}^{-T}\mathbf{L}^{-1}$, it is straightforward to develop a procedure, analogous to the Cholesky decomposition algorithm [40, Section 4.2], to directly compute $\mathbf{L}^{-T}$ rather than through the inverse of $\mathbf{L}$. The embedding matrix $\log(\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}})$ can be written as

$$\underbrace{\log\begin{bmatrix}\mathbf{L}^{-T} & \boldsymbol{\mu}\\ \mathbf{0}^T & 1\end{bmatrix}}_{\boldsymbol{\Sigma}\text{ is full}}\ \text{or}\ \underbrace{\begin{bmatrix}\log\sigma_1 & & & \frac{\mu_1\log\sigma_1}{\sigma_1-1}\\ & \ddots & & \vdots\\ & & \log\sigma_n & \frac{\mu_n\log\sigma_n}{\sigma_n-1}\\ & & & 0\end{bmatrix}}_{\boldsymbol{\Sigma}\text{ is diagonal}}. \qquad (21)$$

Note that for the diagonal covariance $\Sigma=\mathrm{diag}(\sigma_i^2)$, the embedding matrix has a simple, analytic expression. For the full-covariance, performing decomposition of $\mathbf{\Sigma}$ costs $O(n^3/3)$. Though being upper triangular, $\mathbf{A}_{\boldsymbol{\mu},\mathbf{L}^{-T}}$ may not be diagonalizable. We use MATLAB function "logm" to compute the matrix logarithm, which implements the algorithm in [44] with a cost of $O(28(n+1)^3)$. For diagonal-covariance Gaussians, computation of the matrix logarithm is very efficient since it only requires the logarithms, multiplications and divisions of real numbers, each n times.

IE-LogE For indirect embedding based on the left coset, we compute the eigen-decomposition $\begin{bmatrix} \mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^T & \boldsymbol{\mu} \\ \boldsymbol{\mu}^T & 1 \end{bmatrix} = \mathbf{O}\mathrm{diag}(\lambda_i)\mathbf{O}$, where $\lambda_i, i = 1, \ldots, n+1$, are eigenvalues and $\mathbf{O}$ is an orthogonal matrix consisting of eigenvectors corresponding to λ_i. The embedding matrix has the form

$$\log(\mathbf{P}_{\boldsymbol{\mu},\mathbf{L}^{-T}}) = \mathbf{O}\mathrm{diag}(\frac{1}{2}\log(\lambda_i))\mathbf{O}^T. \tag{22}$$

Hence, the complete embedding procedure will cost $O(4(n+1)^3)$ for matrix eigen-decomposition and $O((n+1)^3)$ for matrix multiplications.

For indirect embedding based on the right coset, we first perform matrix factorization $\mathbf{\Sigma} = \mathbf{L}^{-T}\mathbf{L}^{-1}$ whose complexity is $O(n^3/3)$. Then we compute the eigen-decomposition $\begin{bmatrix} \mathbf{L}^{-1}\mathbf{L}^{-T} & \mathbf{L}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\mu}^T\mathbf{L}^{-T} & \boldsymbol{\mu}^T\boldsymbol{\mu}+1 \end{bmatrix} = \mathbf{O}'\mathrm{diag}(\lambda'_i)\mathbf{O}'^T$. The embedding matrix can thus be written as

$$\log(\mathbf{P}'_{\boldsymbol{\mu},\mathbf{L}^{-T}}) = \mathbf{O}'\mathrm{diag}(\frac{1}{2}\log(\lambda'_i))\mathbf{O}'^T. \tag{23}$$

The eigen-decompoistion costs $O(4(n+1)^3)$ and the matrix multiplications that follow in Eq. (23) costs $O((n+1)^3)$.

6 Experimental Evaluation

In this section, we apply the proposed L^{2}EMG to image classification. After introducing the experimental setup, we will make a thorough analysis of parameters on the challenging Pascal VOC 2007 database [45], aiming to achieve an in-depth insight into L^{2}EMG while obtaining a set of suitable parameters before moving to other scenarios. Then, we make comparisons with state-of-the-art local descriptors on Pascal VOC 2007 [45], Caltech-256 [46], Scene-15 [3], Sun-397 [47], and Flickr Material Database (FMD) [48].

6.1 Experimental Setup

We employ the well-known bag-of-visual-words (BoW) pipeline for classification [49], and follow the setting of BoW as described in [50]. Three kinds of encoding methods are considered to model images: (1) hard coding (Vector Quantization, VQ) with sum (average) pooling [49]; (2) Locality-constrained Linear Coding (LLC) with max pooling [51]; and (3) Fisher vector (FV) with average pooling [4]. As suggested in [50], [52], VQ coding vectors are fed to $\mathcal{X}^2$ kernel-based SVM while LLC and FV are fed to linear SVM. We exploit the VLFeat package [53] wherever possible, e.g., extraction of SIFT, LBP or HoG and SVM implementation. We perform power normalization (PN) followed by l_2 normalization [4] for L^{2}EMG and L^{2}ECM descriptors [15]. The spatial information is incorporated via the spatial pyramid matching (SPM) mechanism [3]. Three levels of SPM $[1\times 1, 2\times 2, 3\times 1]$ are used for Pascal VOC 2007, and two levels of SPM $[1\times 1, 3\times 1]$ are used for Caltech-256, Scene-15, and Sun-397. We do not use SPM on FMD.

6.2 Parameter Analysis of L^{2}EMG on VOC 2007

We conduct experiments on VOC 2007 benchmark to make an analysis of parameters involved in L^{2}EMG. VOC 2007 contains 20 classes and 9,963 images in total. We select this dataset for parameter analysis because it is challenging and is well designed [45] so that the conclusions drawn on this dataset can extrapolate to other challenging ones, as argued in [4]. We follow the standard protocol: using "train" and "val" sets for training, "test" set for testing, and mean average precision (mAP) over 20 categories as the accuracy measure. We employ the baseline VQ coding method with 4K dictionary for evaluation, unless otherwise stated.

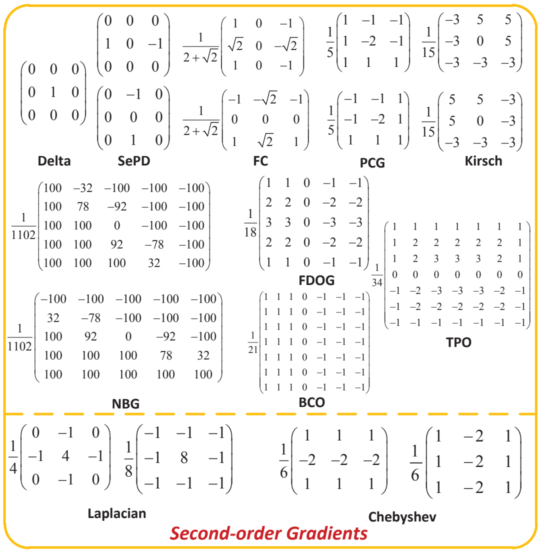


Fig. 2. Image derivative operators or filters [54] used in our paper

TABLE 3
Classification accuracy (mAP, %) of L^{2}EMG versus varying combination of raw features on VOC 2007

No.	Raw features	# Dim.	L^{2}ECM [15]	L^{2}EMG
#1	Delta+Loca.+SePD$_{\mathrm{1st+2nd}}$	7	51.25	47.85
#2	OHE (8 Bins)	8	52.11	51.13
#3	Delta+#2	9	52.83	51.83
#4	Delta+Loca.+#2	11	**52.97**	50.44
#5	FC+PCG+Kirs.+NBG+FDOG+BCO +TPO	16	50.86	52.28
#6	(FC+PCG+Kirs.)$_{\mathrm{1st+2nd}}$+Lapl. +Cheb.	16	50.38	51.32
#7	Delta+#5	17	52.30	53.72
#8	Delta+#6	17	51.87	54.13
#9	Delta+(FC+PCG+Kirs.)$_{\mathrm{1st+2nd}}$ +Lapl.$_{\mathrm{2st+4th}}$	17	51.07	53.93
#10	Delta+(FC+PCG+Kirs.+NBG +FDOG)$_{\mathrm{1st+2nd}}$+Lapl.	23	52.09	**54.32**
#11	Delta+(FC+PCG+Kirs.)$_{\mathrm{1st+2nd+3rd}}$ +(Lapl.+Cheb.)$_{\mathrm{2st+4th}}$	25	51.43	54.29
#12	Delta+Gabor	25	51.40	51.97
#13	RGB+#7	19	51.55	52.28
#14	Lab+Loca.+Harr.+#7	24	50.36	50.04

Raw Features The intention of L^{2}EMG is to leverage local statistics of multiple image cues, what we call raw features. As suggested in [7], the commonly used raw features include intensity, color, location (Loca.), 1st- and 2nd-order derivatives computed by separated pixel difference (SePD) operators (cf. Table 3). Moreover, we study a variety of operators [54], as shown in Fig. 2, which extract image derivatives in varying directions and at different scales. We further consider the orientation histogram of edges, 3rd- and 4th- derivatives, Gabor filters and Harris features. We thus cover most of the commonly used cues in image processing.

We empirically divide the combinations of these cues into four categories containing 14 kinds of raw feature combinations, as presented in Table 3. 1) The first class of raw features (#1 in Table 3) is often used in covariance descriptors [7], which contains intensity, location, and derivatives computed by SePD operators. 2) The orientation histogram of edges (OHE) [55] collects the zero-order statistics of gradients which provides a complementary cue to the aforementioned ones. We combine OHE with intensity or location to form a family of OHE-based raw features (#2 ∼#4 in Table 3). 3) A family of derivative-related image operators are listed as #4∼#11 in Table 3. The operators we considered are 3×3 1st-order derivative operators including Frei-Chen (FC), Prewitt compass gradient (PCG) and Kirsch, 5×5 1st-order Nevatia-Babu gradient (NBG) operators, first-order Derivative operator of Gaussian (FDOG), 7×7 1st-order operators such as Box Car Operator (BCO), Truncated Pyramid Operator (TPO), and 2nd-order operators of Laplacian and Chebyshev. The #5 and #7 raw features focus on combination of the 1st-order operators. The #6, #8 and #10 raw features combine 1st- and 2nd-order operators, while #9 and #11 combine 1st-, 2nd- and 4th-order ones. Note that #5, #7 and #10 raw features integrate multi-scale operators and that the intensity by Delta operator is joined as a basic cue from #7∼#11. 4) Finally, we evaluate additional color, Gabor filters, Harris features and location in #12∼#14.

L^{2}EMG and L^{2}ECM [15] descriptors are both extracted with 16×16 patch size and sampling step 2. Table 3 presents the classification results (mAP, %), from which we can see that their performances against various combinations of raw features are very different. L^{2}ECM achieves the best performance by combining OHE, intensity and location (#4 in Table 3), while L^{2}EMG obtains the highest accuracy by using a combination of multiscale, 1st-, 2nd-order derivative operators and intensity (#10 in Table 3). Generally speaking, L^{2}ECM performs better than L^{2}EMG by using the family of OHI-based features, while L^{2}EMG outperforms L^{2}ECM by using the family of derivative-related image operators.

The choice of raw features in L^{2}EMG or L^{2}ECM influences substantially the classification performance. From the detailed comparison, we have the following conclusions.

1) The family of OHE-based raw features are more appropriate for L^{2}ECM rather than L^{2}EMG. The reason may be that the mean of discrete probability distributions (OHEs) brings a negative effect in embedding Gaussians with full covariances.
2) Comparison of combinations #8, #10 and #11 indicates that the 3rd-order and multi-scale operators slightly improve performance. These three combinations achieve comparable results and hence we do not consider higher-order or larger size operators.
3) From the comparisons of #2 against #3, #5 against #7, and #6 against #8, we can see that intensity is an important cue for both L^{2}ECM and L^{2}EMG descriptors. By integrating it into the descriptor we can achieve 0.8% ∼ 3% performance gains.
4) The Gabor filters are of relatively high dimension but fail to bring better performance. This may be because texture information of small patches extracted by Gabor filters is not distinct enough and is thus not very competing for classification tasks. Color, location and Harris features are not suitable for L^{2}EMG or L^{2}ECM.
5) From combinations #5 to #11, we observe that the inclusion of more raw features improves performance. But when the dimension of feature vector is higher than 25, the performance decreases. Much higher dimensionality may bring difficulties in estimating the full-covariance matrix of a Gaussian, due to small sample size and unaffordable, computational burden [2].

To balance accuracy and efficiency, in all the following experiments, we adopt the combinations of raw features given by #4 and #8 in Table 3 for L^{2}ECM and L^{2}EMG, respectively. The work above should not be simply interpreted as a process of raw features selection, and the raw features are not limited to the ones used here. However, we make the first attempt in analyzing what raw features help to improve the performance of L^{2}EMG; our analysis also underscores that the raw features and their combinations are important, and thus sophisticated feature selection algorithms may further benefit the proposed descriptor.

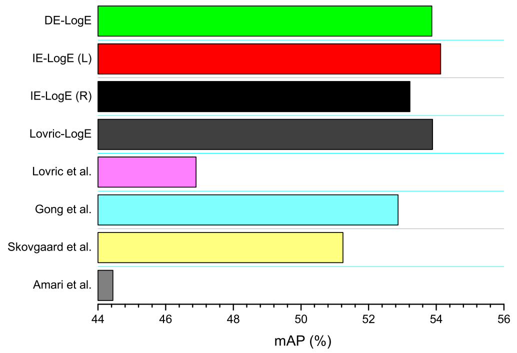


Fig. 3. Comparison of different embedding methods on VOC 2007

Embedding Methods The space of Gaussians is a Riemannian manifold and here we carry out experiments to testify whether our embedding can effectively leverage the geometrical structure of this space. We evaluate the performance of our embedding methods: direct embedding (DE-LogE) as in Eq. (21), indirect embedding based on the left coset (IE-LogE(L)) and based on the right coset (IE-LogE(R)) as in Eq. (22) and Eq. (23), respectively. In the methods of Skovgaard et al. [34], Gong et al. [26], Lovrić et al. [36], Gaussians with the geodesic distance or dissimilarity measure are entangled and thus they are computationally prohibitive in our methodology. Hence, these embedding matrices or vectors are viewed as in the Euclidean space and they are compared with ours as baseline Euclidean

[2] The L^{2}EMG with diagonal-covariance Gaussians are special cases of that with full-covariance Gaussians. Such Gaussians indicate uncorrelations of various raw features which are often violated in practice. Nevertheless, for feature vectors of much higher dimensions, it may be the only feasible solution. In Section 6.3, we will introduce such a competitive L^{2}EMG descriptor with 128-dim SIFT as raw features.

methods. As an exception, the embedding matrices of Lovrić et al. share similar forms to IE-LogE(L), and as in [32], the Log-Euclidean methodology can be applied to them (Lovrić-LogE). We set patch size and sampling step to 16×16 and 2, respectively.

Fig. 3 shows the comparison results. We can see that IE-LogE(L) achieves the highest recognition accuracy, while IE-LogE(R) and DE-LogE outperform all the Euclidean baselines [26], [34], [36]. Finally, we note that Lovrić-LogE [32] is also competitive, and it is a little inferior to IE-LogE(L). We believe that these comparison results validate that our embedding methods respect the geometry of the space of Gaussians. We owe this to that our embedding processes strictly conforms to the Lie group isomorphisms which establish the equivalence between the embedded and embedding spaces.

Patch size & **Sampling step** The patch size and sampling step have influence on capturing local structure and local descriptor density, respectively. We fix the sampling step to 2 to test the effect of patch size, and the results are presented in Table 4(a). The performance increases consistently as patch size gradually reduces from 24×24 to 12×12. This indicates that local characteristics at finer scales are more distinctive and discriminative. But a too small patch size (8×8 or smaller) leads to insufficient number of samples for Gaussian estimation so that performance deteriorates. By combination of four scales of patches (i.e, 12∼24), L^{2}EMG achieves mAP 54.92% and L^{2}ECM 53.78%. In all the remaining experiments, we set patch size to 16×16 whenever single scale L^{2}EMG or L^{2}ECM is used, for ensuring enough sampling points to estimate Gaussians or covariance descriptors.

TABLE 4
Effect (in terms of mAP, %) of patch size and sampling step on L^{2}EMG and L^{2}ECM on VOC2007

(a) Patch size

Patch size	8×8	12×12	16×16	20×20	24×24
L^{2}ECM [15]	52.68	53.12	52.97	51.91	51.57
L^{2}EMG	53.87	54.24	54.13	53.35	52.80

(b) Sampling step

Sampl. step	1	2	3	4	5	6
L^{2}ECM [15]	53.16	52.97	52.36	51.43	50.25	49.10
L^{2}EMG	54.26	54.13	53.49	53.02	52.47	51.58

Next we conduct experiment to test the influence of the sampling step by fixing the patch size to 16×16. It is evident that a smaller step results in denser patches and thus more information is extracted. We vary the sampling step from 1 pixel to 6 pixels. As presented in Table 4(b), with the increase of step size, the classification performances of both descriptors consistently drop. Note that a smaller sampling step means a larger number of image patches and accordingly, higher computational cost for descriptor extraction, coding and pooling. To tradeoff accuracy and efficiency, we set the sampling step to 2 throughout the following experiments.

6.3 Comparison on VOC 2007

Finally, we compare our descriptors with three well-known descriptors, Local Binary Pattern(LBP), HoG and SIFT extracted using VLFeat [53]. The descriptors are compared at single(S) scale and at multiple(M) scales (7 scales for LBP and HoG, and 4 scales for SIFT, L^{2}ECM and L^{2}EMG). To avoid possible comparison bias, we employ three coding methods, i.e., VQ (hard-assignment), LLC (soft-assignment) and FV (super vector).

The parameters of our descriptor, L^{2}EMG (Full), are tuned as described previously. Specifically, we employ 17-dim raw features #8 (cf. Tab. (3)) for estimation of full-covariance Gaussians, which are then embedded by IE-LogE(L). The L^{2}ECM are computed with 11-dim raw features #4. Note that the embedding method of DE-LogE is suitable for high-dimensional features. We propose L^{2}EMG (Diag.) to exploit high-dimensional raw features, i.e., the widely used 128-dim SIFT. We compute pixel-wise SIFT descriptors at four scales, reduce their dimensions to 64 using PCA, and then estimate the diagonal-covariance Gaussians at a single scale on 16×16 patches with sampling step 2 [3]. The diagonal-covariance Gaussians are embedded by DE-LogE (cf. Eq. (21)) to obtain 128-dim L^{2}EMG (Diag.).

Comparison of descriptors using VQ and LLC The comparison of descriptors using VQ and LLC are presented in Table 5. It can be seen that LBP and HoG are much inferior to the other three descriptors in terms of classification accuracy. We observe that in most cases L^{2}ECM has very similar accuracy to SIFT, both of which are, on average, outperformed by L^{2}EMG (Full) (over 1.5% for the case of single scale and over 1.2% for multiple scales). It is worthy to mention that even single scale L^{2}EMG (Full) is superior to multiple scale SIFT with LLC. It can also be observed that, when using VQ and LLC, L^{2}EMG (Diag.) is less competitive than L^{2}EMG (Full).

TABLE 5
Classification accuracy (mAP, %) of competing descriptors using various coding methods on VOC 2007

Methods		VQ		LLC		FV
Dictionary size		4K	25K	4K	25K	256
LBP	S	43.32	44.17	41.54	42.55	48.70
	M	44.27	45.83	42.43	43.95	49.63
HoG	S	43.87	46.14	42.17	43.22	50.34
	M	45.84	48.06	43.25	44.68	52.85
SIFT	S	52.39	53.72	53.42	56.54	59.98
	M	53.82	55.42	53.66	57.66	61.71
L^{2}ECM [15]	S	52.97	53.67	53.44	56.23	57.52
	M	53.78	55.36	53.68	56.89	58.59
L^{2}EMG (Full)	S	54.13	54.67	55.44	58.16	59.66
	M	54.92	55.88	55.87	58.67	60.60
L^{2}EMG (Diag.)	-	52.78	54.46	53.42	55.31	64.65

Comparison of descriptors using FV Unlike VQ and LLC which are both based on K-means clustering (or dictionary learning), FV trains a Gaussian mixture model (GMM) as a dictionary. To handle high-dimensional descriptors, the FV method performs dimension reduction by PCA and exploits Gaussian with diagonal covariance for estimating GMM and deriving the coding vectors. Hence, to begin with, we study the impact of dimension reduction on our descriptors with FV on VOC 2007. The results of L^{2}EMG (Full), L^{2}EMG (Diag.), and L^{2}ECM at single scale are shown in Fig. 4.

[3] Diagonal-covariance Gaussians at multiple scales bring trivial gains which are hence not reported.

It can be seen that they achieve the highest accuracies when dimensions are 96, 80 and 56, respectively. When the dimension > 48, L^{2}EMG (Diag.) has distinct advantages over the other two while L^{2}EMG (Full) outperforms L^{2}ECM.

Subsequently, we compare L^{2}EMG (Full), L^{2}EMG (Diag.), L^{2}ECM, SIFT with reduced dimension, LBP and HoG. For each method, the dimensionality is determined by its highest classification accuracy. The results are presented in the last column of Table 5. The LBP and HoG using the FV coding are still much inferior to other descriptors. For the case of single scale, L^{2}EMG (Full) is comparable to SIFT and is superior to L^{2}ECM (over 2% accuracy); for multi-scale case, SIFT is better than L^{2}EMG (Full) with a margin of 1% accuracy. Finally, it can be observed that L^{2}EMG (Diag.) outperforms all the other descriptors and it is almost 3% higher than the second-best method, SIFT, in accuracy.

Discussion on L^{2}EMG (Full), L^{2}EMG (Diag.) It can be observed that L^{2}EMG (Full) outperforms L^{2}EMG (Diag.) using VQ or LLC by a non-trivial margin but is much inferior to L^{2}EMG (Diag.) using FV. From the perspective of image descriptor, one of the major differences between VQ or LLC and FV is whether PCA is performed or not. In the FV method, the dictionary and coding vector are based on GMM with diagonal covariance matrices; it has been shown [4] that PCA makes the SIFT descriptors better fit the diagonal assumption (the performance is raised over 7% by using PCA). In what follows, we make experiments to investigate whether PCA is suitable for or harms L^{2}EMG (Full).

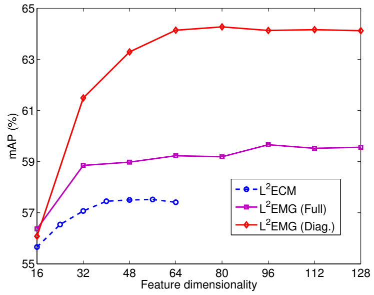


Fig. 4. Impact of dimensionality reduction on L^{2}EMG and L^{2}ECM with FV on VOC 2007

For each of L^{2}EMG (Full) and L^{2}EMG (Diag.), we randomly select 5 million descriptors from training images and estimate per-dimension distribution (histogram). We found that the distributions of all dimensions of L^{2}EMG (Diag.) are unimodal, but over 70 percent of dimensions of L^{2}EMG(Full) are essentially bimodal in distributions. As an example, Fig. 5 shows the distributions of some dimensions of each descriptor. Obviously, L^{2}EMG (Diag.) is more appropriate for diagonal covariance matrix assumption, even without using PCA (indeed L^{2}EMG (Diag.) obtains 63.38% accuracy without PCA as opposed to 64.14% with PCA). In contrast, the distribution of L^{2}EMG (Full) is strongly multi-modal (please refer to the supplementary material for examples). This will degrade much the distinctness of L2EMG(Full).

Comparison with state-of-the-art The proposed L^{2}EMG (Full) and L^{2}EMG (Diag.) are complementary as they employ different raw features. Hence, we also consider the score level fusion of them, in which the fusion weight is determined by using cross-validation. We present the comparison results in Table 6(a). Under the FV framework, L^{2}EMG (Diag.) is much better than SIFT and is even slightly better than the fusion of SIFT and color; fusion of L^{2}EMG (Full) and L^{2}EMG (Diag.) outperforms SIFT+color by 1.2%. Peng et al. [58] improved the VLAD method [13] by leveraging high-order statistics (HOS-VLAD), achieving an accuracy of 61.2%; they further resorted to the supervised dictionary learning and achieved 65.2%, while our fusion method yields 65.7%, still outperforming them. Li et al. [56] combined FV with visual concepts (VC) learned from a large quantity of internet images collected by using text-based queries. Kobayashi [57] transformed the histogram-based features using the Dirichlet-derived GMM Fisher kernel. Our L^{2}EMG (Diag.) and the fusion method both perform better than them.

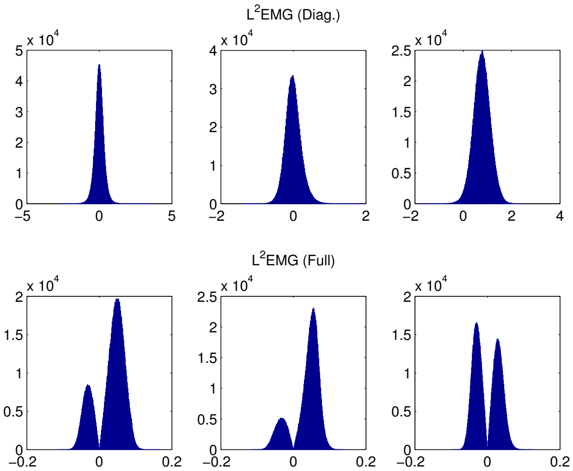


Fig. 5. Comparison of the distributions of dimension #30, #60 and #90 (from left to right) of L^{2}EMG (Diag.) and L^{2}EMG (Full) on VOC 2007

6.4 Results on Other Benchmark Datasets

In what follows, we carry out experiments on other popular benchmark datasets under the FV framework but focus on descriptor comparison. Our experiments involve object classification on Caltech-256 [46], scene categorization on Scene-15 [3] and Sun-397 [47], and material recognition on FMD [48].

Caltech-256 [46] has about 30K images distributed in 256 categories, containing diverse object sizes, poses, and lighting conditions. According to the standard experimental setting, we increase gradually per-category training set size from 15 to 60 with a step of 15 and test with the rest of images. We present the average classification accuracy over five trials in Table 6(d). We compare with the methods in [57], [64], [65], all of which involve mapping or transforming of the local features (e.g., SIFT) and achieve improved performance over SIFT with FV [4]. We also compare with Multipath Hierarchical Matching Pursuit (M-HMP) [66] which learns features in a deep architecture, and yields better performance than fusing SIFT and color with FV. The superiority of L^{2}EMG (Full) over L^{2}ECM

TABLE 6
Comparison with state-of-the-art methods on a variety of benchmark datasets

(a) VOC 2007

Methods	mAP
Challenge winners	59.4
FV+VC [56]	62.9
Kobayashi2014 [57]	63.8
HOS-VLAD [58]	61.2
FV+SIFT [4]	61.8
FV+(SIFT+Color) [4]	63.9
FV+L^2ECM [15]	58.6
FV+L^2EMG(Full)	60.8
FV+L^2EMG(Diag.)	64.7
FV+L^2EMG(Full + Diag.)	65.7

(b) Scene-15

Methods	Acc.
GIST [22]	73.3 ± 0.7
CENTRIST [22]	83.9 ± 0.8
VQ+VC [56]	85.4
Hybrid-Parts + GIST + SP [59]	86.3
CENTRIST + LLC + Boosting [60]	87.8
FV+SIFT [53]	88.1
FV+L^2ECM [15]	85.6 ± 0.3
FV+L^2EMG(Full)	88.1 ± 1.1
FV+L^2EMG(Diag.)	89.5 ± 0.5
FV+L^2EMG(Full + Diag.)	90.2 ± 0.3

(c) FMD

Methods	Acc.
Sharan et al. [61]	57.1
Kobayashi2014 [57]	57.3
DeCAF [62]	60.7±2.0
DTD [63]	61.1±1.4
FV+SIFT [63]	58.2±1.7
FV+(SIFT+Color) [63]	63.3±1.9
FV+L^2ECM [15]	58.2±1.9
FV+L^2EMG (Full)	65.4±1.0
FV+L^2EMG (Diag.)	64.2±1.7
FV+L^2EMG (Full + Diag.)	67.2±0.3

(d) Caltech-256

# training	15	30	45	60
Kernel Map [64]	40.3±0.1	48.5±0.2	52.9±0.3	55.9±0.4
Arandjelovic etal [65]	41.2±0.3	49.5±0.2	53.9±0.4	56.8±0.3
Kobayashi2014 [57]	41.8±0.2	49.8±0.1	54.4±0.3	57.4±0.4
M-HMP [66]	42.7	50.7	54.8	58.0
FV+SIFT [4]	38.5±0.2	47.4±0.1	52.1±0.4	54.8±0.4
FV+(SIFT+Color) [4]	41.0±0.3	49.4±0.2	54.3±0.3	57.3±0.2
FV+L^2ECM [15]	39.4±0.3	47.2±0.4	51.6±0.3	55.0±0.3
FV+L^2EMG(Full)	39.6±0.4	47.6±0.4	52.6±0.2	56.2±0.4
FV+L^2EMG(Diag.)	42.4 ±0.3	50.9±0.2	55.1±0.2	59.0±0.6
FV+L^2EMG(Full + Diag.)	45.0 ±0.2	53.6±0.3	58.2±0.3	61.8±0.4

(e) Sun-397

# training	5	10	20	50
Features Fusion [47]	14.5	20.9	28.1	38.0
Kobayashi2014 [57]	-	-	-	46.1±0.1
DeCAF [62]	-	-	-	40.9±0.3
FV+SIFT [4]	19.2±0.4	26.6±0.4	34.2±0.3	43.3±0.2
FV+(SIFT+Color) [4]	21.1±0.3	29.1±0.3	37.4±0.3	47.2±0.2
FV+L^2ECM [15]	16.2±0.1	23.8±0.3	31.1±0.2	39.9±0.2
FV+L^2EMG(Full)	17.0±0.4	24.7±0.2	32.5±0.2	42.3±0.3
FV+L^2EMG(Diag.)	20.5±0.5	28.6±0.2	36.8±0.2	46.4±0.3
FV+L^2EMG(Full + Diag.)	22.2±0.5	30.8±0.1	39.6±0.1	49.8±0.2

is not significant, and both are comparable to SIFT. Our L^2EMG(Diag.) outperforms all the aforementioned competing methods, even FV combined with SIFT and color; by combining L^2EMG(Full) with L2EMG(Diag.), we further achieve an accuracy improvement of 2.8% on average.

Scene-15 [3] consists of 4,485 images, and the number of images per category varies from 200 to 400. Following [3], we randomly choose 100 training images per class, while the remaining ones are reserved for testing. The average accuracy over five trials is reported. We compare with GIST [67] and CENTRIST [22] which are both designed for scene categorization. From Table 6(b) we can see that L^2EMG(Full) achieves the same accuracy as SIFT, and outperforms GIST, CENTRIST and the learned visual context (VC) features [56]. L^2EMG(Full) performs better than L^2ECM by 2.5%, while L^2EMG(Diag.) is superior to L^2EMG(Full). Combination of L^2EMG(Full) and L^2EMG(Diag.) yields an accuracy of 90.2%.

Sun-397 [47] contains 108,754 images over 397 categories, and each category has at least 100 images. Following the experimental setting in [47], we use ten pre-defined subsets of the dataset for evaluation, where each subset includes 50 training images and 50 testing images per class. In addition, we use different numbers (5, 10, 20 and 50) of images for training, but all the 50 testing images are used for testing in each trial. The results on Sun-397 are reported in Table 6(e). We can see that L^2EMG(Full) performs better than L^2ECM, the feature fusion method [47] and DeCAF [62], but is inferior to SIFT. The reason may be that the current raw features in L^2EMG(Full) may not be very suitable for scene images and dimensionality reduction. The potential raw features appropriate to scene may help improve the performance of L^2EMG(Full). L^2EMG(Diag) performs better than SIFT by 2.25% on average. The fusion of L^2EMG(Diag) and L^2EMG(full) achieves the best performance, outperforming the second best method, SIFT+Color, 1.9% on average.

FMD [48] contains a diverse selection of surfaces in 10 material categories, each including 100 images. According to the evaluation protocol in [63], we report, in Table 6(c), the results averaged over 5 random splits of FMD (50% for training and 50% for testing). We compare with the perceptually inspired features [61] which are specifically designed for material classifications, the deep convolution activation features (DeCAF) [62] and high-level semantics attributes (DTD) [63] learned upon the FV and DeCAF. One can see that L^2EMG(Full), achieving 65.4% in accuracy, is superior to L^2EMG (Diag.) and performs much better than all the competing methods, including the fusion of SIFT and color using FV. The performance is further boosted to 67.5±0.3% by our fusion method, which is slightly higher than the currently best result, 67.1±1.5 %, obtained by a sophisticated method that combines FV+SIFT, DeCAF, DTD learned upon FV and DeCAF [63]. Furthermore, our method has a much smaller standard deviation.

6.5 Experimental Summary and Discussion

This section summarizes and discusses our experimental results based on different types of classification tasks.

1) L^2EMG is a very general image descriptor, suitable for a variety of image classification problems, including object classification, scene categorization and material recognition. It even outperforms some descriptors specifically designed for particular classification tasks.
2) When combined with FV, L^2EMG(Diag) is consistently superior to SIFT, one of the most effective hand-crafted descriptors, and L^2EMG(Full) fits material classification very well. We attribute the efficacy of L^2EMG to the leveraging of local, higher-order statistics, and to the effective embedding methods which respect the algebraic and topological structure of the space of Gaussians.

3) By using VQ and LLC, L^{2}EMG(Full) is superior to L^{2}EMG(Diag.), but is inferior to L^{2}EMG(Diag.) by using FV. The reason may be that PCA degrades its distinctness, as described in Section 6.3. Interestingly, L^{2}EMG(Full) and L^{2}EMG(Diag.) are complementary descriptors, and their combination leads to, in almost every case, state-of-the-art performance.
4) Compared to L^{2}ECM, L^{2}EMG(Full) has better performance on all benchmarks, particularly on VOC 2007, Scene-15 and FMD. This clearly shows that the first-order statistics (mean vector) is by no means trivial.

7 Conclusion

This paper presented a function-valued descriptor called L^{2}EMG to characterize local, high-order statistics by extracting Gaussian distributions from a local neighborhood. We developed Log-Euclidean methods to handle Gaussians with Euclidean operations instead of Riemannian ones. Our main contributions are summarized as follows:

- Unlike popular histogram-based descriptors, which, based on feature space quantization, collect zero-order (occurrence) information, the proposed L^{2}EMG descriptor is continuous and models higher-order statistics. It can naturally leverage multiple cues or other descriptors (e.g, SIFT) as raw features.
- We showed that the space of Gaussians can be equipped with a Lie group structure, and that it is equivalent to a subgroup of the upper triangular matrix group. These conclusions, not presented in previous literature as far as we know, provide new insights into the algebraic and geometrical structure of Gaussians.
- We introduced two novel methods to embed Gaussians in the linear spaces. One performs direct embedding by matrix logarithm while the other performs embedding via the space of SPD matrices as an intermediate process. Both methods depend on Lie group isomorphisms and thus respect the geometry of spaces involved.
- We evaluated thoroughly the L^{2}EMG descriptors, clarifying the influence of raw features, embedding methods and so on. In the BoW pipelines, we compared with a variety of descriptors on popular benchmarks and demonstrated that L^{2}EMG descriptors are very competitive.

Compared to histogram-based descriptors (e.g. SIFT), our L^{2}EMG descriptors are computationally more demanding. In the future, we will develop parallel algorithms running on off-the-shelf, multi-core CPUs or many-core Graphic Processing Units (GPU), to accelerate computation of the descriptors. In light of the success of covariance descriptors, we are also interested to apply L^{2}EMG descriptors to other vision tasks such as visual tracking and image retrieval.

References

[1] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.

[2] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1/2, pp. 43–72, 2005.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2169–2178.

[4] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[5] D. L. Bihan, J. Mangin, C. Poupon, C. Clark, S. Pappata, and N. Molko, "Diffusion tensor imaging: Concepts and applications," *J Magn. Reson. Imaging*, vol. 66, pp. 534–546, 2001.

[6] H. Knutsson, "Representing local structure using tensors," in *Proc. Scan. Conf. on Image Anal.*, 1989, pp. 244–251.

[7] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 589–600.

[8] J. Weichert and H. Hagen, Eds., *Visualization and image processing of tensor fields*. Berlin / Heidelberg: Springer, 2006.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, 2004.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Int. Conf. Comp. Vis. Patt. Recog.*, 2005, pp. 886–893.

[11] J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, July 2010.

[12] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[13] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sept 2012.

[14] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, 2006.

[15] P. Li and Q. Wang, "Local Log-Euclidean covariance matrix (L^{2}ECM) for image representation and its applications," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 469–482.

[16] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. II–506.

[17] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sept 2010.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, pp. 346–359, 2008.

[19] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, 2010.

[20] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1720, 2010.

[21] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution grayscale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[22] J. Wu and J. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug 2011.

[23] R. Gupta and A. Mittal, "SMD: A locally stable monotonic change invariant feature descriptor," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 265–277.

[24] F. Tang, S. H. Lim, N. L. Chang, and H. Tao, "A novel feature descriptor invariant to complex brightness changes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2631–2638.

[25] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *Int. J. Comput. Vision*, pp. 41–66, 2006.

[26] L. Gong, T. Wang, and F. Liu, "Shape of Gaussians as feature descriptors," in *Proc. Int. Conf. Comp. Vis. Patt. Recog.*, 2009, pp. 2366–2371.

[27] H. Nakayama, T. Harada, and Y. Kuniyoshi, "Global Gaussian approach for scene categorization using information geometry," in *Proc. Int. Conf. Comp. Vis. Patt. Recog.*, 2010, pp. 2336–2343.

[28] S. ichi Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford University Press, 2000.

[29] G. Sharma, S. ul Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 1–12.

[30] G. Serra, C. Grana, M. Manfredi, and R. Cucchiara, "Modeling local descriptors with multivariate Gaussians for object and scene recognition," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 709–712.

[31] B. Ma, Q. Li, and H. Chang, "Gaussian descriptor based on local features for person re-identification," in *Workshop on Proc. Asian Conf. Comput. Vis.*, 2014.

[32] P. Li, Q. Wang, and L. Zhang, "A novel Earth Mover's Distance methodology for image matching with Gaussian mixture models," in *Proc. IEEE Int. Conf. Comput.Vis.*, 2013, pp. 1689–1696.

[33] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Math. Soc.*, vol. 37, pp. 81–91, 1945.

[34] L. T. Skovgaard, "A Riemannian geometry of the multivariate normal model," *Scandinavian Journal of Statistics*, vol. 11, no. 4, pp. 211–223, 1984.

[35] M. Calvo and J. M. Oller, "A distance between multivariate normal distributions based on an embedding into the siegel group," *J. Multivar. Anal.*, vol. 35, no. 2, pp. 223–242, 1990.

[36] M. Lovrić and M. Min-Oo, "Multivariate normal distributions parametrized as a Riemannian symmetric space," *J. Multivar. Anal.*, vol. 74, no. 1, pp. 36–48, 2000.

[37] B. Hall, *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Springer, 2003.

[38] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*. Springer-Verlag, 2002.

[39] J. Gallier, "Logarithms and square roots of real matrices," *CoRR*, vol. arXiv:0805.0245, 2013.

[40] D. Kincaid and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing*. American Mathematical Society, 2002.

[41] R. Lopez-Valcarce and S. Dasgupta, "Some properties of the matrix exponential," *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 48, no. 2, pp. 213–215, Feb 2001.

[42] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 1999.

[43] N. J. Higham, "Computing the polar decomposition with applications," *SIAM J. Sci. Stat. Comput.*, vol. 7, no. 4, pp. 1160–1174, 1986.

[44] P. I. Davies and N. J. Higham, "A schur-parlett algorithm for computing matrix functions," *SIAM J. Sci. Comput.*, vol. 25, no. 2, pp. 464–485, 2003.

[45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[46] G. Griffin, A. Holub, and P. Perona, "The Caltech-256," California Institute of Technology, Tech. Rep., 2007.

[47] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010.

[48] L. haran, R. Rosenholtz, and E. H. Adelson, "Material perception: What can you see in a brief glance?" *J. Vis.*, vol. 9, no. 8, p. 784, 2009.

[49] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput.Vis.*, 2003, pp. 1470–1477.

[50] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. Brit. Mach. Vis. Conf*, 2011.

[51] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3360–3367.

[52] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: A comprehensive study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 493–506, March 2014.

[53] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.

[54] W. K. Pratt, *Digital Image Processing, 4th Edition*. New York, NY, USA: John Wiley & Sons, Inc., 2007.

[55] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Int. Workshop on Auto. Face and Gesture Recognit.*, 1995, pp. 296–301.

[56] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 851–858.

[57] T. Kobayashi, "Dirichlet-based histogram feature transform for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3278–3285.

[58] X. Peng, L. Wang, Y. Qiao, and Q. Peng, "Boosting VLAD with supervised dictionary learning and high-order statistics," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 660–674.

[59] Y. Zheng, Y.-G. Jiang, and X. Xue, "Learning hybrid part filters for scene recognition," in *Europ. Conf. on Comp. Vis.*, 2012.

[60] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *Proc. Int. Conf. Comp. Vis. Patt. Recog.*, 2011, pp. 2777–2784.

[61] L. Sharan, C. Liu, R. Rosenholtz, and E. H. Adelson, "Recognizing materials using perceptually inspired features," *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 348–371, 2013.

[62] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

[63] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3606–3613.

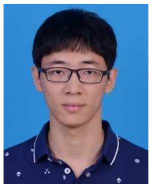
[64] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, 2012.

[65] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2911–2918.

[66] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2013.

[67] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

Peihua Li is a professor in School of Information and Communication Engineering, Dalian University of Technology. He received the PhD degree from Harbin Institute of Technology in 2002, and was awarded the honorary nomination of National Excellent Doctoral dissertation in 2005. He was supported by Program for New Century Excellent Talents in University of Ministry of Education of China in 2011. Currently he is mainly interested in image classification and search using theoretical and computational methods of information geometry. He has published over fifty papers in referred conferences and journals.



Qilong Wang is a PhD candidate in School of Information and Communication Engineering, Dalian University of Technology. His research interests include image and video classification and recognition. He has published several papers in top conferences including ICCV, CVPR and ECCV.

Hui Zeng is pursuing his Msc degree in School of Information and Communication Engineering, Dalian University of Technology. His research interests include image retrieval and search. As a leading member, he achieved the second-best ranking among 843 teams in Alibaba Large-scale Image Search Challenge.

Lei Zhang (M'04, SM'14) received his Ph.D degree from Northwestern Polytechnical University, China, in 2001. He joined the Dept. of Computing, The Hong Kong Polytechnic University, as an Assistant Professor in 2006. Since July 2015, he has been a Full Professor in the same department. His research interests include Computer Vision, Pattern Recognition, Image and Video Processing, and Biometrics, etc. Dr. Zhang has published more than 200 papers in those areas. By 2015, his publications have been cited more than 15,000 times in literature. He is currently an Associate Editor of IEEE TIP, IEEE TCSVT and Image and Vision Computing. He was selected as the Highly Cited Researcher by Thomson Reuters, 2015.