Image classification based on log-Euclidean Fisher Vectors for covariance matrix descriptors

Sara Akodad, Lionel Bombrun, Charles Yaacoub, Yannick Berthoumieu, Christian Germain

Image classification based on log-Euclidean Fisher Vectors for covariance matrix descriptors

Sara Akodad[1], Lionel Bombrun[1], Charles Yaacoub[2], Yannick Berthoumieu[1] and Christian Germain[1]
[1] Laboratoire IMS, Université de Bordeaux, CNRS, UMR 5218, 351 cours de la libération, 33400 Talence, France
e-mail: sara.akodad@ims-bordeaux.fr, lionel.bombrun@ims-bordeaux.fr, yannick.berthoumieu@ims-bordeaux.fr, christian.germain@ims-bordeaux.fr
[2] Faculty of Engineering, Holy Spirit University of Kaslik (USEK), P.O. Box 446 Jounieh, Lebanon
e-mail: charlesyaacoub@usek.edu.lb

***Abstract*— This paper introduces an image classification method based on the encoding of a set of covariance matrices. This encoding relies on Fisher vectors adapted to the log-Euclidean metric: the log-Euclidean Fisher vectors (LE FV). This approach is next extended to full local Gaussian descriptors composed by a set of local mean vectors and local covariance matrices. For that, the local Gaussian model is transformed to a zero-mean Gaussian model with an augmented covariance matrix. All these approaches are used to encode handcrafted or deep learning features. Finally, they are applied in a remote sensing application on the UC Merced dataset which consists in classifying land cover images. A sensitivity analysis is carried out to evaluate the potential of the proposed LE FV.**

***Keywords*— Fisher vector, vector of locally aggregated descriptors, log-Euclidean metric, covariance matrices, SIFT descriptors, deep neural network, classification.**

I. Introduction

The goal of a supervised classification algorithm is to label an image with one class name depending on its content. The leading approaches in the beginning of the 2000s were based on feature coding. These approaches include the bag of words model (BoW) [1], the vector of locally aggregated descriptors (VLAD) [2], [3] and the Fisher vectors (FV) [4], [5], [6]. All these approaches have been successfully validated in a wide variety of applications such as image classification [4], [7], [8], text retrieval [9], action and face recognition [10], etc.

Since 2012, Convolutional Neural Networks (CNN) have become a standard for image classification problems [11], [12]. Since then, in order to take advantage of both approaches (deep neural network architecture and FV descriptors), many authors have proposed hybrid classification algorithms which combine them. For example, a network of fully connected layers has been trained on the FV descriptors in [13]. The deep Fisher network composed by stacking several FV layers is another hybrid architecture [14]. Inspired by the VLAD image representation, the NetVLAD has been proposed in [15] to mimick a VLAD layer. Other approaches include the FV or VLAD encoding of CNN features from different layers of the network [16], [17], [18], [19].

Recently, many works are dedicated to extend the formalism of encoding to features lying in a non-Euclidean space. This is for example the case of covariance matrices that have already proved their interest in many classification problems [10], [20], [21], [22]. But, since covariance matrices are symmetric positive definite (SPD) matrices, standard Euclidean calculus are not adapted. The use of the geometrical information can lead to more accurate representation of the inherent structure of these observed covariance matrices. Since then, the Riemannian geometry has become increasingly popular in the computer vision community. When dealing with covariance matrices, two Riemannian metrics are generally considered: the log-Euclidean and the affine-invariant Riemannian metric. Recently, some authors have proposed to use these metrics in order to extend the BoW and VLAD descriptors. This yields to the so-called log-Euclidean bag of words (LE BoW) [23], [24], bag of Riemannian words (BoRW) [25], log-Euclidean vector of locally aggregated descriptors (LE VLAD) [10] and intrinsic Riemannian vector of locally aggregated descriptors (RVLAD) [10]. Recently, we have proposed to extend the FV descriptors to SPD features. This has involved the Log-Euclidean Fisher vectors (LE FV) and the Riemannian Fisher vectors (RFV) [26], [27], [28]. Although the log-Euclidean metric do not yield full affine invariance compared to the affine invariant Riemannian metric, it is invariant by similarity (orthogonal transformation and scaling). These characteristics allow to reduce the computation time while maintaining comparable performance with methods based on the affine invariant Riemannian metric. Nevertheless, all these coding methods are only valid for covariance matrix descriptors. Here, we propose to extend these methods to full local Gaussian descriptors that are composed of local means and local covariance matrices.

The main contributions of the paper are threefold. First, we present how FV can be used to encode a set of covariance matrix by using the log-Euclidean metric. Next, we extend this encoding to full local Gaussian descriptors. The proposed approach relies on an augmented SPD matrix which represents both the local means and the covariance matrices [29], [30], [31]. And finally, we propose an hybrid deep-learning architecture which allows to encode CNN features with the proposed LE FV.

The paper is structured as follows. Section II recalls the general principle of a classification algorithm based on FV. Section III introduces the proposed LE FV descriptor and its adaptation to full local Gaussian descriptors. An application of

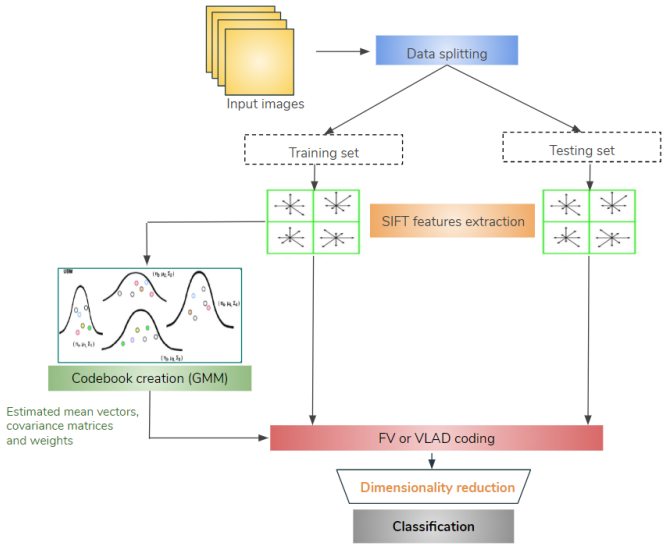

Fig. 1. Workflow of the general process : SIFT features coding for image classification.

these descriptors to the classification of remote sensing images is presented in Section IV. Section V introduces an hybrid deep learning architecture based on the encoding of CNN layers with the LE FV. And finally, Section VI synthesizes the main conclusions of this work.

II. VLAD and Fisher vectors encoding of SIFT descriptors

In this section, we present the very conventional approach which consists in encoding handcrafted features such as scale invariant feature transform (SIFT) descriptors [32] with the Fisher vectors (FV) [4], [5], [6] or with the vector of locally aggregated descriptors (VLAD) [2]. This approach has been successfully used in a wide variety of image processing applications [4], [7], [8], [9], [10]. The global principle of this approach is presented in Fig. 1. It consists in the following steps:

1) *Data splitting*: At the beginning, data are separated into two disjoints sets used respectively for training and testing.
2) *Feature extraction*: An handcrafted features extraction step is then applied, with for example SIFT descriptors. This latter allows to obtain feature vectors invariant to scale, rotation, translation, and partially invariant to illumination. The SIFT algorithm is achieved through two stages. First, a detection algorithm is applied to detect keypoints in the image. Then, the SIFT descriptor $\mathbf{x}$ is computed for each detected keypoint. In the end, each image is represented by a set of N SIFT descriptors $\mathbf{x}_i$ of dimension 128.
3) *Codebook creation*: A codebook (also named dictionary) is generated with the previously extracted descriptors computed on the training set. It consists in extracting a set of visual words which are classically obtained by a clustering algorithms, such as k-means or expectation-maximization (EM) algorithms. Formally, a Gaussian mixture model (GMM) is assumed to model the set $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) \in \mathbb{R}^{T\times d}$ of d-dimensional local features extracted from the training images (with $d = 128$ for the SIFT descriptors). The probability density function of the GMM model is given by:

$$p(\mathbf{x}_t|\lambda) = \sum_{i=1}^{K} \omega_k p_k(\mathbf{x}_t|\lambda), \tag{1}$$

where, for each cluster k:

$$p_k(\mathbf{x}_t|\lambda) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}_t-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}_t-\mu_k)\}}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}, \tag{2}$$

where $(\cdot)^T$ is the transpose operator, $|\cdot|$ is the determinant, $\omega_k \in (0,1)$, $\mu_k \in \mathbb{R}^d$, $\Sigma_k \in \mathcal{P}_d$ the space of $d\times d$ symmetric positive definite matrices. In addition, the covariance matrix is assumed to be diagonal, *i.e.* $\sigma_k^2 = \mathrm{diag}(\Sigma_k) \in \mathbb{R}^d$ is the variance vector. For a GMM model, a k-means or EM algorithm can be employed to estimate the parameters of each mixture component (μ_k, σ_k^2 and ω_k). These elements represent a codeword and the set composed by the K codewords gives the codebook. The k-means (resp. the EM) algorithm is used when the SIFT features are encoded with the VLAD (resp. the FV) descriptors.
4) *Fisher vector and VLAD encoding*: Let $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ be a set of d-dimensional vectors extracted form an image. The FV or VLAD encoding consists in projecting these descriptors in the previously learned codebook.

The FV descriptor associated to $\boldsymbol{\mathcal{X}}$ is obtained as the gradient of the sample log-likelihood with respect to the parameters of the GMM model, scaled by the inverse square root of the Fisher information matrix (FIM) $\mathbf{F}_\lambda$:

$$\mathscr{G}_\lambda^{\boldsymbol{\mathcal{X}}} = \mathbf{F}_\lambda^{-\frac{1}{2}} \nabla_\lambda \log p(\boldsymbol{\mathcal{X}}|\lambda). \tag{3}$$

Next, by deriving with respect to the mean, the dispersion or the weight parameters, the following three FV are obtained:

$$\mathscr{G}_{\mu_k^d}^{\boldsymbol{\mathcal{X}}} = \frac{1}{\sqrt{\omega_k}} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)\left(\frac{\mathbf{x}_n^d - \mu_k^d}{\sigma_k^d}\right), \tag{4}$$

$$\mathscr{G}_{\sigma_k^d}^{\boldsymbol{\mathcal{X}}} = \frac{1}{\sqrt{2\omega_k}} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)\left(\frac{\left[\mathbf{x}_n^d - \mu_k^d\right]^2}{\left(\sigma_k^d\right)^2} - 1\right), \tag{5}$$

$$\mathscr{G}_{\omega_k}^{\boldsymbol{\mathcal{X}}} = \frac{1}{\sqrt{\omega_k}} \sum_{n=1}^{N} \Big(\gamma_k(\mathbf{x}_n) - \omega_k\Big), \tag{6}$$

where μ_k^d (resp. σ_k^d) is the d^{th} element of vector μ_k (resp. σ_k) and $\gamma_k(\mathbf{x}_n)$ is the occupancy probability of $\mathbf{x_n}$ to the k-th Gaussian component, also named the posterior probability, and is defined as:

$$\gamma_k(\mathbf{x}_n) = \frac{\omega_k p_k(\mathbf{x}_n|\lambda_k)}{\sum_{j=1}^{K} \omega_j p_j(\mathbf{x}_n|\lambda_j)}. \tag{7}$$

In the following, only the two gradients with respect to the mean μ_k^d and standard deviation σ_k^d are considered since the state-of-the-art in computer vision have shown that the best results are obtained with these FV [4], [5], [6].

The VLAD descriptor has been introduced in [2] in a similar spirit to the FV. It can be interpreted as an hard version of the FV where only the derivative with respect to the mean is considered in (3). The homoscedasticity assumption should also be made *i.e.* $\sigma_k = \sigma$. To summarize, the VLAD encoding of $\boldsymbol{\mathcal{X}}$ is obtained as the concatenation of vectors $\mathbf{v}_k$:

$$\mathbf{VLAD} = [\mathbf{v}_1^T, \ldots, \mathbf{v}_K^T], \tag{8}$$

where, for each atom of the codebook, the vector $\mathbf{v}_k$ contains the sum of differences between the codeword and the feature samples assigned to it:

$$\mathbf{v}_k = \sum_{\mathbf{x}_n \in c_k} \mathbf{x}_n - \mu_k. \tag{9}$$

Once the FV or VLAD descriptors are computed, a post-processing step is classically employed to improve the classification performance [5], [8]. It consists on a power and an ℓ_2 normalization.

5) *Dimension reduction*: Since the dimensionality of the feature space (FV or VLAD descriptors) can be high, a dimension reduction step can be considered to avoid the curse of dimensionality phenomenon. For that, different unsupervised or supervised dimension reduction techniques can be employed such as principal component analysis (PCA), linear discriminant analysis (LDA), Kernel discriminant analysis (KDA) to cite a few of them.
6) *Classification*: This final step consists on making a decision for each test image based on information contained in their vector representation (FV or VLAD). In practice, various classifiers can be employed such as k-nearest neighbors, support vector machine (SVM) or random forest.

III. Log-Euclidean Fisher Vectors for covariance matrices encoding

The objective of this section is to explain how the classification algorithm detailed in Section II can be adapted to local covariance matrix descriptors. Indeed, since covariance matrices are positive definite matrices, conventional tools developed in the Euclidean space are not well adapted to these observations. The characteristics of the Riemannian geometry of the space $\mathcal{P}_d$ of $d \times d$ symmetric and positive definite (SPD) matrices should be considered in order to obtain appropriate algorithms. Here, we address this point by considering the log-Euclidean metric.

A. Log-Euclidean Fisher Vectors (LE FV)

First, the covariance matrices of handcrafted features (such as SIFT descriptors calculated on a regular grid) are computed on a sliding window. It yields that each image is represented by a set $\boldsymbol{\mathcal{M}} = \{\mathbf{M}_n\}_{n=1:N}$ of covariance matrices $\mathbf{M}_n \in \mathcal{P}_d$. Employing the log-Euclidean metric to analyze these covariance matrices is equivalent to use the Euclidean metric on the log-Euclidean space. For that, each covariance matrix $\mathbf{M}_n$ is first mapped on the log-Euclidean space by applying the matrix logarithm $\mathbf{M}_n^{LE} = \log \mathbf{M}_n$ [24], [33], [34]. Next, a vectorization operator is applied to obtain the log-Euclidean vector representation. To summarize, the log-Euclidean vector representation of an SPD matrix $\mathbf{M}$ is the vector $\mathbf{m} \in \mathbb{R}^{\frac{d(d+1)}{2}}$ defined as $\mathbf{m} = \text{Vec}(\log(\mathbf{M}))$ where Vec is the vectorization operator defined as:

$$\text{Vec}(\mathbf{X}) = \left[X_{11}, \sqrt{2}X_{12}, \ldots, \sqrt{2}X_{1m}, X_{22}, \sqrt{2}X_{23}, \ldots, X_{mm}\right], \tag{10}$$

with X_{ij} the elements of $\mathbf{X}$. Now that SPD matrices are mapped on the log-Euclidean metric space, all the algorithms developed on the Euclidean space can be employed, in particular the FV and VLAD encoding which yield respectively to the so-called log-Euclidean FV (LE FV) [27] and log-Euclidean VLAD (LE VLAD) [10] descriptors.

Note that since dense SIFT descriptors may contain redundant information, the covariance matrices of these descriptors will not be well conditioned. To circumvent this and in order to reduce the dimension of the log-Euclidean vector representation $\mathbf{m}$, a dimension reduction step such as PCA is applied on the handcrafted features as a pre-processing step. This step allows also to better fit the diagonal covariance matrix assumption made in Section II when deriving the FV, *i.e.* $\sigma_k^2 = \text{diag}(\Sigma_k)$. In the following, N_{PCA} will refer to the number of retained principal components.

B. LE FV for full local Gaussian descriptors

In the previous subsection, only the local covariance matrix descriptor has been considered. In a full local Gaussian descriptor, the local mean vector can be jointly exploited with the local covariance matrix in order to increase the image representation within the classification task. Based on the works of [29], the local Gaussian model can be transformed to a zero-mean Gaussian model with an augmented SPD matrix of dimension $(d+1) \times (d+1)$ given by:

$$\mathbf{M}_{augmented} = |\mathbf{M}|^{-\frac{1}{d+1}} \begin{bmatrix} \mathbf{M} + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix}. \tag{11}$$

This approach has been successfully validated by many authors for image classification tasks [30], [31]. Now that local mean and covariance matrix are embedded in a larger SPD matrix, the LE FV and LE VLAD descriptors can be computed for full local Gaussian descriptors by following the same strategy as the one described in Section III.

IV. Application to remote sensing image classification

This section introduces an application to remote sensing image classification. For that, the UC Merced land use land

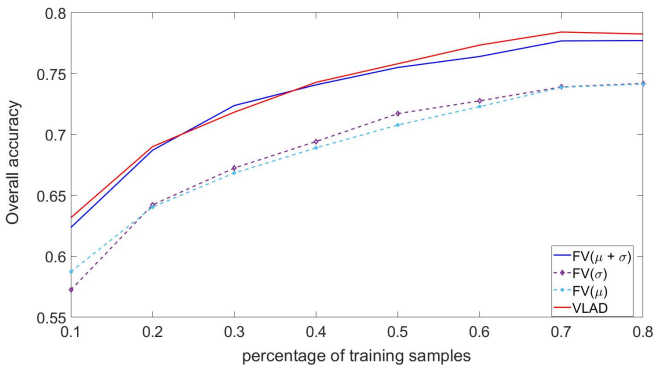

Fig. 2. Classification accuracy for the FV and VLAD encoding methods.

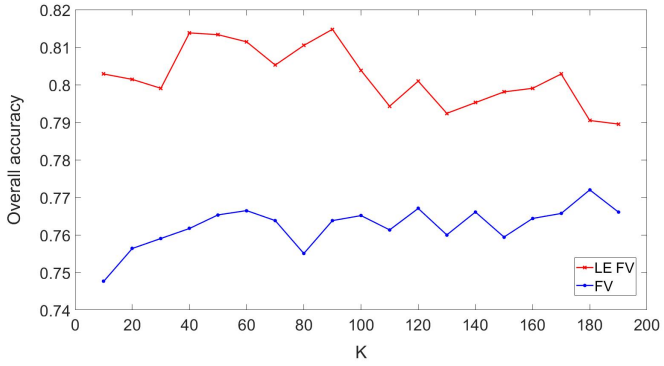

Fig. 3. Influence of the codebook dimension on the FV and LE FV classification accuracy.

cover dataset is considered [35]. It contains 21 land cover classes such as beach, golf course, harbor with 100 color images per class of dimension 256×256 pixels with a spatial resolution of 1 foot. The aim of the following experiments is to compare the different classification methods presented in Sections II and III.

In the following, a linear SVM classifier is considered for the final step in Fig. 1. The performance is evaluated in term of overall accuracy where half of the database is used for training while the remaining half is used for testing.

A. Comparison between FV and VLAD encoding

Fig. 2 draws the evolution of classification accuracy for the FV (in blue) and VLAD (in red) based encoding methods as a function of the proportion of training samples for $K = 30$. In addition, three versions of FV descriptors are considered, *i.e.* by considering separately the derivative with respect to the mean μ_k^d, the dispersion σ_k^d or by fusing them, noted respectively $FV(\mu)$, $FV(\sigma)$ and $FV(\mu+\sigma)$. One can observe that best performances are obtained for the VLAD and FV (when both derivatives are considered) descriptors with a similar behavior. In the following, only FV descriptor will be used.

B. Comparison between FV and LE FV

The purpose of this second experiment is to illustrate the potential of the proposed LE FV descriptor. For that, Fig. 3 draws the evolution of the classification accuracy for the FV (in blue) and the LE FV (in red). As observed, the best results are observed for the proposed LE FV illustrating the interest of local covariance matrix descriptor. A significant gain of about 6% if observed for the LE FV compared to FV. Note also that it does not require a large codebook, performances are quite stable with the codebook dimension.

In Fig. 4, we evaluate the influence of the number of retained principal components N_{PCA} in the pre-processing step for $K = 30$. As observed, the dimension reduction greatly impacts the classification accuracy. It is hence necessary to find an optimal trade-off between performance and resources

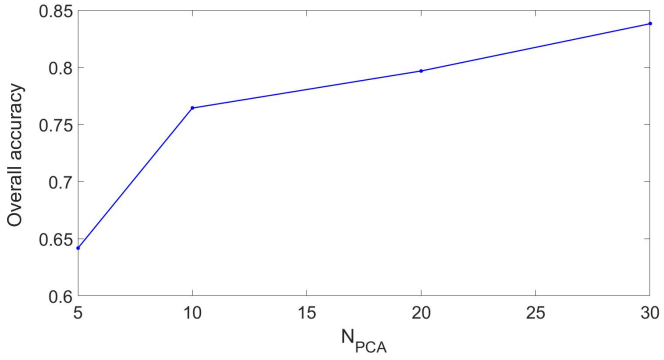

Fig. 4. Influence of the number of retained principal components on classification accuracy for LE FV.

Table 1. Comparison between FV and LE FV descriptors for $K = 30$ and $N_{PCA} = 20$

Descriptor	Overall accuracy $\pm$ standard deviation
FV [4]	75.5 $\pm$ 1.3%
LE FV	79.4 $\pm$ 1.3%
LE FV augmented	80.1 $\pm$ 1.2%

consumption in terms of computational time and memory space.

As explained in Section III-B, the local mean vector can be exploited jointly with the local covariance matrix to form an augmented SPD matrix. In order to evaluate the interest of this approach, Table 1 summarizes the classification performance obtained on the UC Merced dataset for the FV, LE FV and LE FV computed on the augmented SPD matrices for $K = 30$ and $N_{PCA} = 20$. As observed the best classification result is obtained with these augmented descriptors illustrating the interest of a full local Gaussian descriptor.

V. Log-Euclidean Fisher Vectors in an Hybrid Deep Learning Architecture

A. FV encoding of CNN features

Now that the interest of the proposed LE FV has been observed for the classification of handcrafted features such

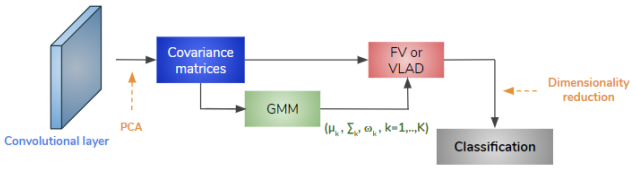

Fig. 5. LE FV encoding for CNN features

Table 2. Classification results for the first and second convolutional layers of vgg-verydeep-16 model

Descriptor	Conv 1	Conv 2
FV [4]	58.2 ± 0.7%	62.5 ± 1.2%
LE FV	70.3 ± 1.0%	87.7 ± 0.7%
LE FV augmented	66.3 ± 1.1%	84.8 ± 0.2%

as SIFT, an application to deep learning is introduced. Convolutional Neural Networks (CNNs) have proven to be very effective in many image processing applications such as classification, segmentation and synthesis. Recently, some hybrid architectures have been proposed to combine FV and CNN in order to benefit from both approaches. For example, in [19], Li *et al.* have proposed an hybrid architecture which extract FV computed on the ouput of different convolutional layers of a pre-trained neural network. This strategy has shown competitive results for remote sensing image classification. The main idea consists in using multi-layer features of a pre-trained CNN model. Once the features are extracted, improved FV encoding is applied to generate features vectors. Those vectors are then fused with features of fully-connected layers to represent the mid-level feature vectors of a scene image. Finally, a linear SVM classifier is applied for classification purpose.

Inspired by this approach, we propose to encode the output of the convolutional layers with the LE FV defined in Section III. Fig. 5 introduces the workflow of this proposed approach for a single convolutionnal layer. As a preliminary approach, we propose to encode only the first and second layers of a CNN. Indeed, for the deepest layers, the spatial dimension of the output are too small to extract a set of covariance matrices. The LE FV encoding can hence only be done for the first layers. Table 2 synthesizes the classification results obtained on the first (Conv 1) and second (Conv 2) layers of the vgg-verydeep-16 model [36]. As observed, a gain is recorded for the proposed LE FV compared to the conventional FV approach. This is in agreement with the previous experiment on SIFT features where a similar conclusion has been done when encoding handcrafted features. Nevertheless, note that for this approach, the exploitation of the mean vector in the augmented SPD matrix (LE FV augmented) seems useless.

B. Global hybrid deep learning architecture

Now that LE FV have shown promising results to encode the first layers of a CNN. We propose an hybrid architecture which combines these LE FV with FV computed on the last layers of the CNN. The general framework is described in Fig. 6. Table 3 summarizes the classification results obtained for the LE FV encoding of SIFT and CNN descriptors.

Table 3. Classification comparison between the two approaches comprising the LE FV encoding

Descriptor	Overall accuracy ± standard deviation
SIFT	79.4 ± 1.3%
CNN	94.4 ± 0.1%

As observed on this experiment on the UC Merced database, the combination of LE FV computed on the first two layers and FV computed on the last five CNN layers as shown in Fig. 6 allows to improve the classification performance (94.4 ± 0.1%) and exceeds the classification results obtained with the SIFT descriptors (79.4 ± 1.3%). A significant gain of 15% is observed.

VI. Conclusion

In this paper, an image classification algorithm has been introduced. It consists in encoding a set of local covariance matrices with Fisher vectors (FV) derived with the log-Euclidean metric: the log-Euclidean Fisher vectors (LE FV). An extension of this method to full local Gaussian descriptors has been proposed. It consists in computing the LE FV on augmented symmetric positive definite matrices gathering both the local mean vectors and local covariance matrices. The experiment on the UC Merced land use land cover dataset have shown the potential of the proposed approach compared to FV for the encoding of handcrafted features (such as SIFT) or convolutional neural networks features.

References

[1] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Oct 2005, vol. 1, pp. 370–377 Vol. 1.

[2] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[3] R. Arandjelović and A. Zisserman, "All about VLAD," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[4] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[5] F. Perronnin, J. Sánchez, and T. Mensink, *Improving the Fisher kernel for large-scale image classification*, vol. 6314 of *Lecture Notes in Computer Science*, pp. 143–156, Springer Berlin Heidelberg, 2010.

[6] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010*, 2010, pp. 3384–3391.

[7] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and Fisher vectors for efficient image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 745–752.

[8] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, 2013.

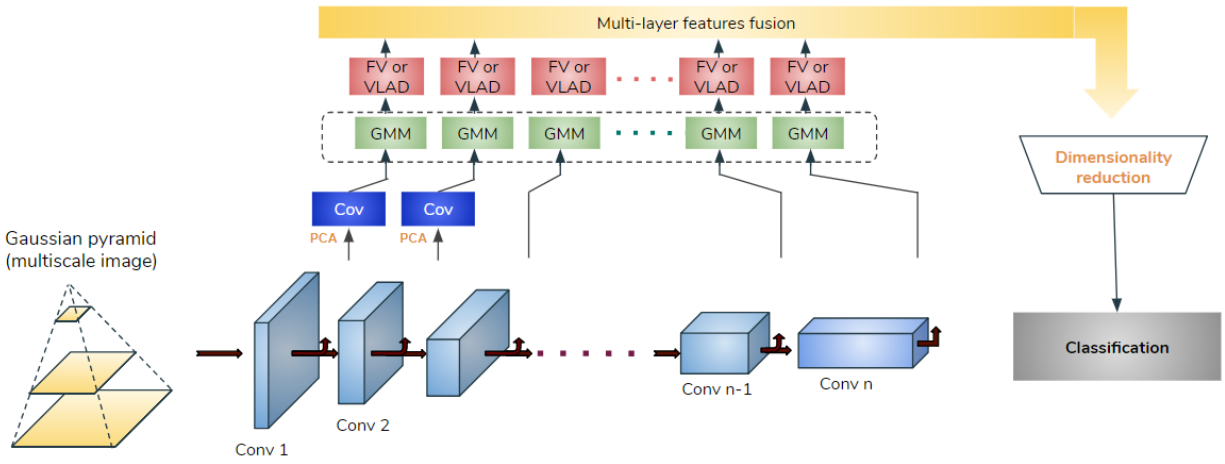

Fig. 6. Hybrid deep learning architecture based on encoding covariance matrices derived from CNN layers with LE FV

[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[10] M. Faraki, M. T. Harandi, and F. Porikli, "More about VLAD: A leap from Euclidean to Riemannian manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4951–4960.

[11] Y. Le Cun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzky, Ed., pp. 396–404. Morgan-Kaufmann, 1990.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, USA, 2012, NIPS'12, pp. 1097–1105, Curran Associates Inc.

[13] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3743–3752.

[14] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep fisher networks for large-scale image classification," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, USA, 2013, NIPS'13, pp. 163–171, Curran Associates Inc.

[15] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: CNN architecture for weakly supervised place recognition," *CoRR*, vol. abs/1511.07247, 2015.

[16] J. Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," *CoRR*, vol. abs/1504.05133, 2015.

[17] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 65–94, May 2016.

[18] A. Diba, A. M. Pazandeh, and L. Van Gool, "Deep visual words: Improved fisher vector for image classification," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, May 2017, pp. 186–189.

[19] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, Oct 2017.

[20] P. Formont, F. Pascal, G. Vasile, J. Ovarlez, and L. Ferro-Famil, "Statistical classification for heterogeneous polarimetric SAR images," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 567–576, 2011.

[21] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *NeuroComputing*, vol. 112, pp. 172–178, 2013.

[22] S. Said, L. Bombrun, and Y. Berthoumieu, "Texture classification using Rao's distance on the space of covariance matrices," in *Geometric Science of Information*, 2015.

[23] C. Yuan, W. Hu, X. Li, S. Maybank, and G. Luo, *Human action recognition under log-Euclidean Riemannian metric*, pp. 343–353, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[24] M. Faraki, M. Palhang, and C. Sanderson, "Log-euclidean bag of words for human action recognition," *IET Computer Vision*, vol. 9, no. 3, pp. 331–339, 2015.

[25] M. Faraki, M. T. Harandi, A. Wiliem, and B. C. Lovell, "Fisher tensors for classifying human epithelial cells," *Pattern Recognition*, vol. 47, no. 7, pp. 2348 – 2359, 2014.

[26] I. Ilea, L. Bombrun, C. Germain, R. Terebes, M. Borda, and Y. Berthoumieu, "Texture image classification with Riemannian Fisher vectors," in *IEEE International Conference on Image Processing*, 2016, pp. 3543 – 3547.

[27] I. Ilea, L. Bombrun, S. Said, and Y. Berthoumieu, "Covariance matrices encoding based on the log-Euclidean and affine invariant Riemannian metrics," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018, CVPRW'18.

[28] I. Ilea, L. Bombrun, S. Said, and Y. Berthoumieu, "Fisher vector coding for covariance matrix descriptors based on the log-Euclidean and affine invariant Riemannian metrics," *Journal of Imaging*, vol. 4, no. 7, 2018.

[29] M. Lovric, M. Min-Oo, and E. A Ruh, "Multivariate normal distributions parametrized as a riemannian symmetric space," *Journal of Multivariate Analysis*, vol. 74, no. 1, pp. 36 – 48, 2000.

[30] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen, "Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 720–729.

[31] M. T. Harandi, M. Salzmann, and R. I. Hartley, "Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods.," *CoRR*, vol. abs/1605.06182, 2016.

[32] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[33] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors," in *Magnetic Resonance in Medicine*, Aug 2006, vol. 56, pp. 411–421.

[34] R. Rosu, M. Donias, L. Bombrun, S. Said, O. Regniers, and J. P. Da Costa, "Structure tensor Riemannian statistical models for CBIR and classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 248–260, Jan 2017.

[35] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2010, GIS '10, pp. 270–279, ACM.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.