An ensemble learning approach for the classification of remote sensing scenes based on covariance pooling of CNN features

Sara Akodad*, Solène Vilfroy*, Lionel Bombrun*, Charles C. Cavalcante†, Christian Germain* and Yannick Berthoumieu*

* Université de Bordeaux, CNRS, IMS, UMR 5218, Groupe Signal et Image, F-33405 Talence, France
e-mail: {sara.akodad, lionel.bombrun, christian.germain, yannick.berthoumieu}@ims-bordeaux.fr
† Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza-CE 60440-900, Brazil
e-mail: charles@ufc.br

***Abstract*—This paper aims at presenting a novel ensemble learning approach based on the concept of covariance pooling of CNN features issued from a pretrained model. Starting from a supervised classification algorithm, named multilayer stacked covariance pooling (MSCP), which exploits simultaneously second order statistics and deep learning features, we propose an alternative strategy which employs an ensemble learning approach among the stacked convolutional feature maps. The aggregation of multiple learning algorithm decisions, produced by different stacked subsets, permits to obtain a better predictive classification performance. An application for the classification of large scale remote sensing images is next proposed. The experimental results, conducted on two challenging datasets, namely UC Merced and AID datasets, improve the classification accuracy while maintaining a low computation time. This confirms, besides the interest of exploiting second order statistics, the benefit of adopting an ensemble learning approach.**

***Index Terms*—Covariance pooling, pretrained CNN models, multilayer feature maps, ensemble learning approach, remote sensing scene classification.**

I. Introduction

A supervised classification algorithm aims at labelling an image to a class according to its content. To this end, standard approaches were based on encoding handcrafted features with for instance the bag of words model (BoW) [1], the vector of locally aggregated descriptors (VLAD) [2] or the Fisher vectors (FV) [3]. Those coding methods have demonstrated successful results in a large variety of applications such as image classification [3]–[5], text retrieval, action and face recognition, etc. More recently, deep learning methods have proved to outperform standard machine learning algorithms in a large variety of domains. In particular, neural networks models the human brain by stacking multiple layers able to extract and learn automatically specific image features. Convolutional Neural Networks (CNN) [6] have achieved great success in the computer vision community which makes them a standard for image classification tasks [6]. They are built from various hidden layers consisting in convolution, pooling and activation functions. In order to take advantage of coding methods and CNN features, several authors have proposed hybrid architectures which combine CNN models with FV and VLAD descriptors such as the Fisher network [7] and the NetVLAD [8]. A multilayer approach has recently been proposed in [9]. It consists in stacking the FV descriptors computed on the outputs of different CNN layers. Nevertheless, all these strategies operate only with first order statistics in the feature space of the CNN architecture. They do not consider second order statistics which have been proved to be important in human visual recognition process [10].

To tackle this problem, many authors have proposed to define strong and discriminant feature representation by considering second-order statistics with the use of covariance matrices. But, since the geometry of covariance matrices is not Euclidean, standard Euclidean tools are not suited to handle these kind of descriptors. For that, the geometry of the space of symmetric positive definite (SPD) matrices should be taken into account. Since then, on one side, coding methods were extended to covariance matrix descriptors yielding to the following approaches: the log-Euclidean bag of words (LE BoW), the bag of Riemannian words (BoRW) [11], the log-Euclidean vector of locally aggregated descriptors (LE VLAD) [12] and the intrinsic Riemannian vector of locally aggregated descriptors (RVLAD), the Log-Euclidean Fisher vectors (LE FV) [13] and the Riemannian Fisher vectors (RFV) [14]. On the other side, second order statistics were also extended to CNN models to enhance their performance. Several architectures have recently been introduced. A first attempt was the pooled covariance matrix from the outputs of a CNN [15]. Another way to exploit second-order statistics in a deep neural network is the Riemannian SPD matrix network (SPDNet) [16]. This network aims at mimicking the conventional fully connected, convolution-like layers and rectified linear units (ReLU)-like layers of a CNN to data which do not lie on an Euclidean space. In the same spirit, Yu *et al.* have proposed in [17] a second order CNN (SO-CNN) that can be trained in an end-to-end fashion. Recently, we have proposed in [13] an hybrid deep neural network based on the log-Euclidean Fisher vectors encoding of region

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

covariance matrices which combines second-order statistics with FV descriptors. Later, He *et al.* have proposed in [18] a multiscale version: the multilayer stacked covariance pooling (MSCP). Inspired by this work and by the success of ensemble learning strategies in the computer vision community [19], this paper aims at proposing a novel ensemble learning approach based on covariance pooling of stacked convolutional layers.

The paper is structured as follows. Section II presents and discusses the related works based on second order statistics. Then, Section III introduces the proposed ensemble learning approach based on covariance pooling (ELCP). An application on remote sensing scene classification is next presented in Section IV. Finally, Section V concludes this paper and provides some perspectives of this work.

II. Related works

The success achieved by CNN modeling for many computer vision tasks such as image classification has allowed to extend their use in the remote sensing community. However, training a CNN model from scratch requires a large number of labelled images. In computer vision, the ImageNet dataset is generally considered but, when working with remote sensing images, there is no such large annotated dataset. As such, a transfer learning approach is better suited. It consists in exploiting CNN models (pretrained on ImageNet) as feature extractors. This strategy has been proved to be effective and to outperform traditional handcrafted feature-based methods [6]. Based on this concept of transfer learning, different supervised classification strategies have recently been introduced in the literature in order to exploit second order statistics on the output of a CNN:

- LE FV encoding of CNN layers (Hybrid LE FV) [13], [20],
- Multilayer stacked covariance pooling (MSCP) [21].

The next two subsections present these strategies.

A. LE FV encoding of CNN layers

We have introduced in [20] an hybrid deep neural network based on the log-Euclidean Fisher vectors encoding of region covariance matrices. This approach generalizes the algorithm introduced in [9] by exploiting second-order statistics, via the computation of region covariance matrices. The general principle can be summarized as follows. Each layer of a CNN is represented by a set of region covariance matrices which are further encoded with FV. For that, the concept of FV encoding has been extended to covariance matrix descriptors which are SPD matrices. In order to manipulate these data points that do not lie on an Euclidean space but on a Riemannian manifold, the log-Euclidean metric was adopted. This allows the definition of the log-Euclidean Fisher vectors (LE FV). The proposed approach was then integrated in a supervised image classification algorithm based on the encoding of deep neural networks features obtained via a transfer learning approach.

Note that in this approach, only the first layers of the CNN have been encoded with the LE FV since the last convolutional layers have a relatively small spatial dimension. It is hence irrelevant to compute a large number of region covariance matrices for the deepest convolutional layers. A simple encoding with FV is considered (*i.e.* only first order statistics are considered as in [9]). Even if this approach was successfully validated for the classification of large scale images and very high resolution texture images, it involves an heavy computation time. To alleviate this issue, a covariance pooling based approach has been introduced in [21].

B. Multilayer stacked covariance pooling

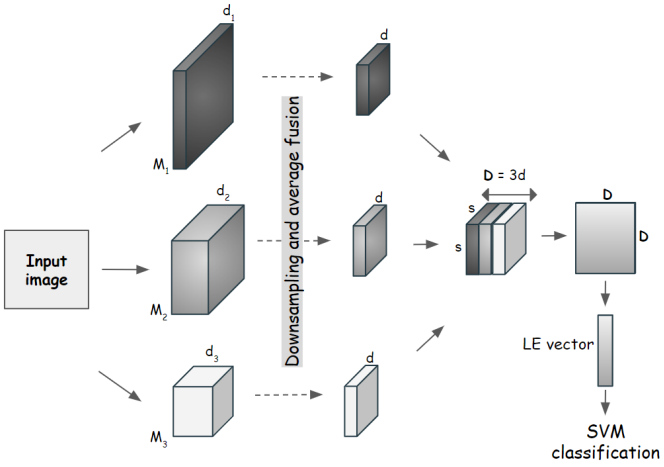


Fig. 1: Architecture of the multilayer stacked covariance pooling strategy (MSCP).

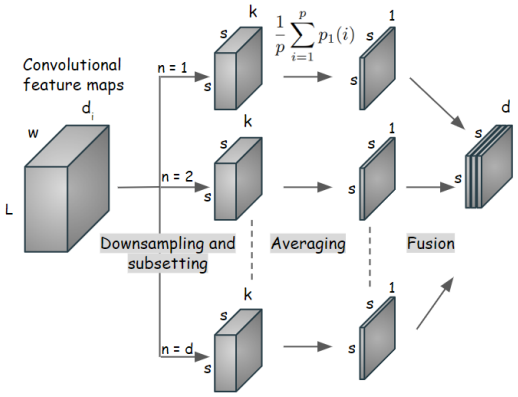


Fig. 2: Description of the downsampling and averaging fusion operations over convolutional feature maps in the MSCP algorithm.

Willing to exploits second order statistics and convolutional networks, He *et al.* have proposed in [21] a method named multilayer stacked covariance pooling (MSCP) where information among specific convolutional layers are fused in a single covariance matrix. Contrary to the hybrid LE FV method presented in Section II-A which represents each layer by a set of covariance matrices, a single covariance matrix is computed for MSCP which allows to significantly faster the computation time.

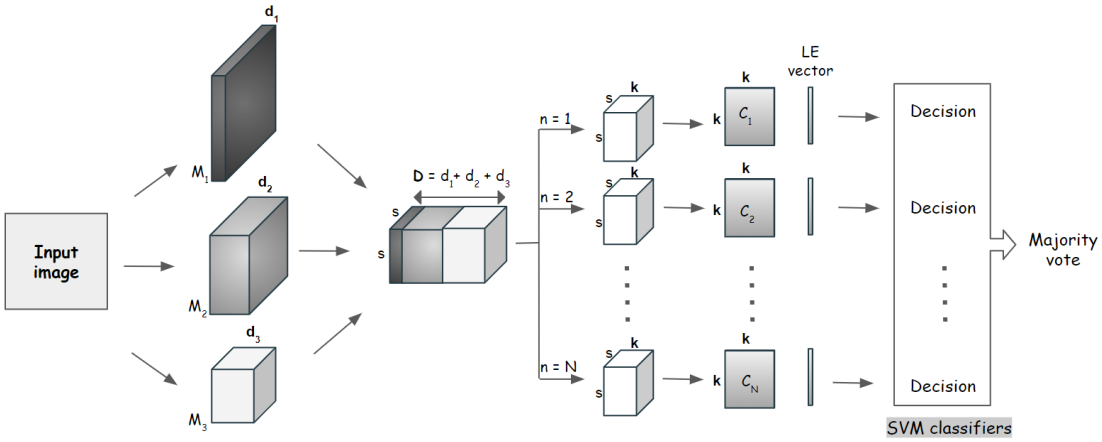


Fig. 3: Architecture of the proposed ensemble learning approach for covariance pooling of multilayer stacked CNN features (ELCP).

The general principle of MSCP is summarized in Fig. 1. First, three convolutional layers, with different depth, are considered and analyzed separately. For a given convolutional layer, an ensemble learning approach is considered by splitting the convolutional features into d subsets. For each subset, k features are selected without replacement. These latter are next downsampled and averaged in order to obtain only one descriptor by subset (see Fig. 2). Then, the d average descriptors for each convolutional layer are concatenated allowing to obtain a tensor of dimension $s \times s \times 3d$. The covariance pooling operator is next applied, it consists in computing the $3d \times 3d$ covariance matrix descriptor. Finally, the log-Euclidean metric is considered for classification. For that, the LE vector representation is computed and an SVM classifier is adopted.

Even if an ensemble learning strategy is adopted in the MSCP algorithm, each convolution map is only seen once. There is no sampling with replacement as classically done in a random forest classifier for example. Moreover, the main drawback of MSCP concerns the averaging operator presented in Fig. 2. There is no reason for the obtained average descriptor of the first subset to be different from the one computed on the last subset. This yields that the covariance matrix of these average descriptors may not be well conditioned. To overcome these issues, the next section introduces a novel ensemble learning approach based on the covariance pooling of CNN features.

III. Ensemble learning approach based on covariance pooling (ELCP)

Inspired by the MSCP classification method presented in Section II-B, the proposed approach, named ELCP, aims at proposing an ensemble learning approach. The idea behind this technique is to combine several weak classifiers to produce better predictive performance than with a single classifier. This will ensure robustness by combining the decision obtained on different subsets [19]. The global principle is shown in Fig. 3.

First, features from three convolutional layers ($conv_1$, $conv_2$ and $conv_3$) are extracted. Their associated feature maps are denoted M_1, M_2 and M_3 respectively. Usually, the spatial dimension of CNN layers is different from one layer to another. For example, in the case of the Vgg-vd-16 model, $M_1 \in \mathbb{R}^{56\times56\times256}$, $M_2 \in \mathbb{R}^{28\times28\times512}$ and $M_3 \in \mathbb{R}^{14\times14\times512}$. In order to stack these three feature maps, a bilinear interpolation is applied allowing to downsample each feature maps to the smallest spatial dimension. Furthermore, for each image, the stacked feature maps produced by the convolutional layers are partitioned into N subsets of k features. These subsets are obtained by random sampling with replacement. Then, for each subset, a covariance pooling strategy is considered. It consists in computing the $k \times k$ covariance matrix $\mathbf{C}$. Since covariance matrices are SPD matrices that lies on a Riemannian manifold, the Euclidean distance is not adapted to compute a similarity measure between them. The geometry of SPD matrices should be considered. Here, the log-Euclidean metric is adopted. It permits the projection of covariance matrices to a tangent space where Euclidean tools can be considered. It allows the representation of covariance matrices as vectors. Practically, it consists in mapping the covariance matrices $\mathbf{C}$ on the log-Euclidean space via the log map operator [22] defined as:

$$\mathbf{C}^{LE} = \log_{\mathbf{I}_d} \mathbf{C} = \mathbf{V}\ \mathrm{logm}(\mathbf{D})\ \mathbf{V}^T, \tag{1}$$

where $\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ is the eigen decomposition of covariance matrix $\mathbf{C}$ and $\mathrm{logm}()$ is the matrix logarithm. Then, to obtain the log-Euclidean vector representation, a vectorization operation $\mathrm{Vec}()$ is performed such that:

$$\mathbf{x} = \mathrm{Vec}(\mathbf{X}) = \left[X_{11}, \sqrt{2}X_{12}, \ldots, \sqrt{2}X_{1k}, X_{22}, \sqrt{2}X_{23}, \ldots, X_{kk}\right]. \tag{2}$$

More detailed explanations about the covariance mapping into the log-Euclidean space can be found in [20]. It yields that the obtained covariance matrix $\mathbf{C} \in \mathbb{R}^{k\times k}$ is transformed to a vector $\mathbf{c} \in \mathbb{R}^{\frac{k(k+1)}{2}}$. Then for each subset, these vectors are fed to a base classifier allowing to obtain a decision. In the end,

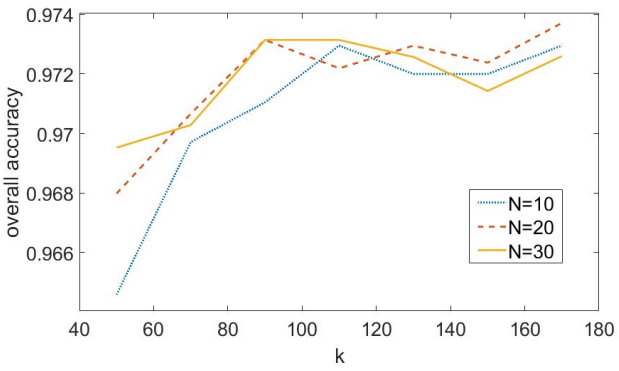
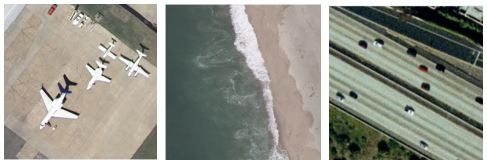


Fig. 4: Influence of the number N of subsets and the number k of selected features in each subset on the classification accuracy.



(a) UC Merced Land Use dataset

(b) AID dataset

Fig. 5: Samples from the datasets used in the classification experiments.

a majority vote is performed to select the most represented decision among the N subsets.

As explained, two parameters should be tuned: the number N of subsets and the number k of selected features in each subset. In order to evaluate the sensitivity of the proposed ELCP approach, an experiment is conducted. Fig. 4 draws the evolution of the classification accuracy as a function of k for different values of N ($N = 10$ to $N = 30$). As observed, the best results are obtained when $N = 20$ and $k = 170$. These parameters are set to these values in the following.

IV. Experiments

This section introduces an application to large scale scene remote sensing image classification. For that, two challenging dataset are considered to evaluate the performance of the proposed supervised classification algorithm: the UC Merced Land Use and the AID datasets. The first one is composed of 21 classes (*e.g.* forest, beach, sparse residential, etc) where each class contains 100 images of dimension 256×256 pixels. The second one, named AID dataset, contains 10 000 aerial images of dimension 600×600 pixels partitioned into 30 classes. Fig. 5 illustrates some images from each dataset. For each dataset, images are randomly separated into training and testing sets according to a fixed rate. 50 % of images are retained for training for the UC Merced dataset while for AID dataset, due to its large size, only 10 % of samples are selected for training. In the following, two CNN models pretrained on ImageNet are considered: AlexNet [6] and Vgg-vd-16 [23]. Note also that the final classification step in Fig. 3 is performed by the linear SVM classifier.

A. Classification results for a single convolutional layer

In this part, the proposed ELCP approach is tested when CNN feature maps are issued from a single layer. Some comparisons are carried out with two other strategies: (1) an hybrid architecture based on the FV encoding of CNN features (Hybrid FV) [9] and (2) the MSCP algorithm [21] detailed in Section II-B. Table I summarizes the classification results obtained on the UC Merced dataset for three convolutional layers for AlexNet and Vgg-vd-16 CNN models.

TABLE I: Classification performance obtained on UC Merced dataset using Hybrid FV, MSCP and the proposed ELCP approaches.

	AlexNet		
	$Conv_3$	$Conv_4$	$Conv_5$
Hybrid FV [9]	92.5 $\pm$ 0.2 %	93.9 $\pm$ 0.3 %	94.1 $\pm$ 0.7 %
MSCP [21]	93.7 $\pm$ 0.2 %	94.6 $\pm$ 0.6 %	93.6 $\pm$ 0.7 %
ELCP	**95.1** $\pm$ **0.4** %	**95.4** $\pm$ **0.4** %	**95.2** $\pm$ **0.4** %
	Vgg-vd-16		
	$Conv_{3,3}$	$Conv_{4,3}$	$Conv_{5,3}$
Hybrid FV [9]	91.8 $\pm$ 0.5 %	96.1 $\pm$ 0.6 %	94.1 $\pm$ 0.4 %
MSCP [21]	87.6 $\pm$ 1.1 %	94.6 $\pm$ 0.6 %	**95.1** $\pm$ **0.2** %
ELCP	**95.4** $\pm$ **0.6** %	**97.0** $\pm$ **0.3** %	**95.1** $\pm$ **0.5** %

As observed in Table I, the proposed ELCP architecture allows to improve the classification accuracy compared to Hybrid FV and MSCP architectures when a single layer is considered. A mean average gain of about 1.5 % and 2.6 % are respectively observed for AlexNet and Vgg-vd-16 models. Note also that the best results are obtained for the Vgg-vd-16 model. In the following, only this CNN model will be considered.

B. Classification results for multilayer features

Now that the proposed ELCP approach has successfully been validated for a single layer, the potential of a multilayer version is investigated. Table II shows the classification results in terms of overall accuracy (mean $\pm$ sd) obtained on the UC Merced and AID datasets for five strategies. The first two ones (CNN and Hybrid FV [9]) exploit only first order statistics computed on the CNN feature maps. The former consists on SVM classifier applied on the fully connected features of the Vgg-vd-16 model pretrained on ImageNet. The two multilayer architectures presented in Section II, namely Hybrid LE FV [20] and MSCP [21], are also compared with the proposed ELCP approach.

TABLE II: Classification performance obtained on UC Merced (50 %) and AID (10 %) datasets using CNN, Hybrid FV, Hybrid LE FV, MSCP and the proposed ELCP approaches

Method	UC Merced	AID
CNN	84.2 ± 2.8 %	76.2 ± 0.4 %
Hybrid FV [9]	96.2 ± 0.7 %	85.8 ± 0.1 %
Hybrid LE FV [20]	96.7 ± 0.2 %	87.6 ± 0.1 %
MSCP [21]	96.7 ± 0.3 %	87.9 ± 0.2 %
ELCP	**97.4 ± 0.3** %	**88.5 ± 0.4** %

As observed in Table II, the use of second order statistics allows to increase the classification performance. On the two datasets, a significant gain of 0.7% is observed for the proposed ELCP approach compared to the best state-of-the-art strategy. This clearly illustrates the benefit of exploiting jointly an ensemble learning strategy and second order statistics in the feature space of a CNN model.

V. Conclusion and perspectives

This paper has introduced a novel supervised classification algorithm based on second-order statistics (*i.e.* covariance matrix descriptors) computed on the output of a deep neural network. Inspired from the principle of ensemble learning techniques, it consists in splitting the CNN features maps into k subsets selected randomly with replacement. Then, a covariance pooling strategy is adopted allowing the modeling of second order statistics. Next, by exploiting the log-Euclidean (LE) metric, these covariance matrices are represented by their LE vector representation that are fed to an SVM classifier. In the end, a majority vote is done to obtain the final decision. The proposed approach has been validated for an application in remote sensing scene classification. For that, two challenging datasets have been considered: the UC Merced and AID datasets. Some comparison with two state-of-the-art algorithms, namely hybrid LE FV and MSCP, have proved the potential of the proposed method. It allows to obtain competitive classification performances while maintaining a reasonable computation time. Future works may include the proposition of an ensemble learning technique which will exploit jointly first and second-order statistics.

Acknowledgment

The authors would like to thank Prof. D. Newsam and G. Xia for providing the UC Merced and the AID datasets. This work was supported by the CNES and by the GROSS project from CNRS-CONFAP. The authors would also acknowledge the financial support of Bordeaux Science Agro and the regional council of Nouvelle Aquitaine.

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*, vol. 1, Oct 2005, pp. 370–377 Vol. 1.

[2] R. Arandjelović and A. Zisserman, "All about VLAD," in *IEEE CVPR*, 2013.

[3] 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.

[4] 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.

[5] 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.

[6] 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*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[7] 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*, ser. NIPS'13. USA: Curran Associates Inc., 2013, pp. 163–171.

[8] 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. [Online]. Available: http://arxiv.org/abs/1511.07247

[9] 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.

[10] B. Julesz, E. N. Gilbert, and J. D. Victor, "Visual discrimination of textures with identical third-order statistics," *Biological Cybernetics*, vol. 31, no. 3, pp. 137–140, 1978.

[11] 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.

[12] 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.

[13] S. Akodad, L. Bombrun, C. Yaacoub, Y. Berthoumieu, and C. Germain, "Image classification based on log-Euclidean Fisher vectors for covariance matrix descriptors," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Xi-an, China, Nov. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01930156

[14] 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*, ser. CVPRW'18, 2018.

[15] C. Ionescu, O. Vantzos, and C. Sminchisescu, "Matrix backpropagation for deep networks with structured layers," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2965–2973.

[16] Z. Huang and L. V. Gool, "A Riemannian network for SPD matrix learning," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2036–2042.

[17] K. Yu and M. Salzmann, "Second-order convolutional neural networks," *CoRR*, vol. abs/1703.06817, 2017. [Online]. Available: http://arxiv.org/abs/1703.06817

[18] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool, "Covariance pooling for facial expression recognition," *CoRR*, vol. abs/1805.04855, 2018. [Online]. Available: http://arxiv.org/abs/1805.04855

[19] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, May 2003.

[20] S. Akodad, L. Bombrun, J. Xia, Y. Berthoumieu, and C. Germain, "Hybrid deep neural network based on the log-Euclidean Fisher vectors encoding of region covariance matrices," *IEEE Transactions on Geoscience and Remote Sensing*, submitted, 2019.

[21] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 6899–6910, Dec 2018.

[22] 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*, vol. 56, no. 2, Aug 2006, pp. 411–421.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556