Encodage de matrices de covariance par les vecteurs de Fisher log-euclidien : application à la classification supervisée d'images satellitaires

Sara Akodad, Lionel Bombrun, Yannick Berthoumieu, Christian Germain

Université de Bordeaux, CNRS, IMS, UMR 5218, Groupe Signal et Image, F-33405 Talence, France.
{sara.akodad, lionel.bombrun, yannick.berthoumieu, christian.germain}@ims-bordeaux.fr

Résumé − Cet article présente une nouvelle architecture hybride basée sur l'encodage par vecteurs de Fisher (VF) des sorties des couches convolutives d'un réseau de neurones. L'originalité de ce travail repose sur l'exploitation des statistiques d'ordre deux via le calcul des matrices de covariance locales. Considérant les propriétés intrinsèques à la géométrie Riemannienne propre à l'espace des matrices de covariance, nous proposons d'utiliser la métrique log-euclidienne afin d'étendre le concept des VF pour l'encodage de matrices de covariance : les vecteurs de Fisher log-euclidiens (LE VF). L'architecture proposée est ensuite évaluée sur différentes bases de données de télédétection : la base UC Merced Land Use Land Cover, la base AID, ainsi que sur deux jeux de données Pléiades sur des forêts de pins maritimes et de parcs ostréicoles.

Abstract − This paper introduces a new hybrid architecture based on Fisher vector encoding (VF) of the convolutional layer outputs of a neural network. The originality of this work is based on the exploitation of second-order statistics via the calculation of local covariance matrices. Considering the intrinsic properties of the Riemannian manifold of covariance matrices, we propose to use the log-euclidean metric in order to extend the concept of VF encoding: the log-euclidean Fisher vectors (LE VF). The proposed architecture is then evaluated on different remote sensing databases : the UC Merced Land Use Land Cover database, the AID database, as well as on two Pléiades datasets on maritime pine forests and oyster beds.

1 Introduction

Les algorithmes de classification supervisée ont pour objectif l'attribution d'une image à une classe particulière selon son contenu. Au début des années 2000, les approches de l'état de l'art étaient basées sur l'encodage des descripteurs locaux par des méthodes telles que les sacs de mots (BoW) [1], les vecteurs de Fisher (VF) [2] ou encore les vecteurs VLAD [3]. Récemment, devant le succès grandissant des réseaux de neurones convolutifs (CNN) [4] pour des tâches de classification, plusieurs travaux se sont orientés vers des architectures hybrides. C'est le cas notamment du réseau de Fisher [5] et de l'architecture NetVLAD [6] qui tirent profit des avantages qu'offre les réseaux de neurones et les méthodes d'encodage. Néanmoins, ces stratégies n'exploitent pas les statistiques de second ordre qui ont démontré d'excellents résultats dans divers problèmes de classification [7]. Pour cela, de nombreux auteurs ont dédié leurs travaux à l'exploitation des statistiques de second ordre afin de définir des représentations compactes et discriminantes de descripteurs. En particulier, les méthodes d'encodage ont été étendues au cas de descripteurs de type matrices de covariance. Ces dernières étant des matrices symétriques définies positives (SPD), il est nécessaire de prendre en compte leur géométrie afin d'adapter les outils classiques de la géométrie euclidienne à ces descripteurs. Pour cela, diverses métriques peuvent être considérée comme la métrique log-euclidienne et la métrique affine-invariante. De cette façon, les méthodes d'encodage classiques ont été entendues aux cas des matrices de covariance : les sacs de mots log-euclidien (LE BoW), les sacs de mots Riemannien (BoRW), les vecteurs VLAD log-euclidien (LE VLAD) et Riemannien (RVLAD) ou encore les vecteurs de Fisher log-euclidien (LE VF) [8, 9] et Riemannien (RVF). Parallèlement à ces travaux, différents auteurs ont proposé des réseaux de neurones basés sur le calcul de matrices de covariance entre les sorties de couches convolutives. C'est le cas du réseau SPDNet proposé dans [10] ou encore du *second-order CNN* (SO-CNN) [11] qui effectue un entraînement de bout en bout du réseau. Récemment, une approche multi-échelles appelée MSCP a été proposée dans [12].

Afin de tirer partie des réseaux CNN, des statistiques de second ordre et des méthodes de codage par VF, ce papier propose une architecture hybride basée sur l'encodage LE VF des matrices de covariance issues des sorties des couches convolutives d'un CNN. L'article est structuré comme suit. La section 2 rappelle le principe général de l'architecture proposée dans [13] basé sur l'encodage par vecteurs de Fisher des sorties des couches convolutives. Puis nous généralisons cette approche dans la partie 3 afin d'exploiter les statistiques de second ordre. Afin d'illustrer l'efficacité de l'approche sur des images à grande échelle

ainsi que sur des images texturées, diverses expérimentations sont réalisées dans la partie 4. Enfin, la partie 5 résume les conclusions et perspectives de ce travail.

2 Encodage par VF des sorties des couches convolutives

En se basant sur le succès des architectures hybrides citées dans l'introduction, Li *et al.* [13] ont proposé une structure hybride permettant d'encoder par vecteurs de Fisher (VF) les sorties de chaque couche convolutives d'un CNN. A l'issue de chaque couche de convolution, les échantillons d'apprentissage sont partitionnés dans l'espace des descripteurs à l'aide du modèle de mélange de gaussiennes (GMM). Pour un ensemble $(\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_T) \in \mathbb{R}^{T\times d}$ de descripteurs d'apprentissage, de dimension d, la densité de probabilité du modèle GMM est définie par :

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

tel que, pour chaque cluster k :

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

où $(\cdot)^T$ est l'opérateur transposé, $|\cdot|$ est le déterminant, $\omega_k \in [0,1]$, $\mu_k \in \mathbb{R}^d$ et $\Sigma_k \in \mathcal{P}_d$ l'espace des matrices $d\times d$ symétriques définies positives. De plus, l'hypothèse classique de matrice de covariance diagonale est supposée (*i.e.* $\sigma_k^2 = \mathrm{diag}(\Sigma_k) \in \mathbb{R}^d$ est le vecteur de variances [2]). Afin d'estimer les paramètres de chaque composante k, à savoir (μ_k, σ_k^2 and ω_k), un algorithme EM est employé. L'ensemble de ces paramètres estimés pour les K composantes du modèle forme le dictionnaire.

Ensuite, les descripteurs sont encodés par l'intermédiaire des VF. Le principe de ces derniers est d'encoder une image par le gradient de la log-vraisemblance par rapport aux paramètres du modèle prédéfini normalisé par l'inverse de la racine carrée de la matrice d'information de Fisher $\mathbf{F}_\lambda$. Soit $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_N)$ l'ensemble de N descripteurs de dimension d issus d'une image. Le VF associé à $\boldsymbol{\mathcal{X}}$ s'écrit :

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

Les dérivées par rapport à la moyenne μ_k^d et l'écart type σ_k^d conduisent aux deux VF suivants :

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

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

avec μ_k^d (resp. σ_k^d) est le $d^{\text{ème}}$ élément du vecteur μ_k (resp. σ_k) et $\gamma_k(\mathbf{x}_n)$ la probabilité d'appartenance de $\mathbf{x}_n$ à la $k^{\text{ème}}$ gaussienne, que l'on nomme également la probabilité a posteriori, définie par :

$$\gamma_k(\mathbf{x}_n) = \frac{\omega_k p_k(\mathbf{x}_n|\lambda_k)}{\sum_{j=1}^{K}\omega_j p_j(\mathbf{x}_n|\lambda_j)}. \tag{6}$$

Finalement, ces VF sont utilisés dans un algorithme de classification par SVM linéaire afin de prédire les classes d'appartenance des échantillons de test.

3 Extension aux descripteurs de type matrices de covariance

Dans la littérature, l'information modélisée par les matrices de covariance a démontré d'excellents résultats, notamment dans les domaines de la reconnaissance faciale, la détection ou encore la classification d'images. Cette section permet de décrire la méthodologie proposée afin d'intégrer les statistiques d'ordre 2 dans le processus de classification présenté dans la partie 2.

3.1 La métrique log-euclidienne

Les matrices de covariance, étant des matrices symétriques définies positives, sont caractérisées par une géométrie Riemannienne spécifique. Arsigny *et al.* ont proposé dans [14] d'exploiter la métrique log-euclidienne. Elle permet de bénéficier des outils classiques de la géométrie euclidienne après avoir projeté les matrices de covariance sur le plan tangent à un point de référence $\mathbf{M}_{ref}$. Cette projection est obtenue par l'opérateur *log map* défini par :

$$\mathbf{M}_n^{LE} = \log_{\mathbf{M}_{ref}} \mathbf{M}_n \tag{7}$$

$$= \mathbf{M}_{ref}^{\frac{1}{2}} \log\left(\mathbf{M}_{ref}^{-\frac{1}{2}}\mathbf{M}_n\mathbf{M}_{ref}^{-\frac{1}{2}}\right)\mathbf{M}_{ref}^{\frac{1}{2}}. \tag{8}$$

L'opération de vectorisation Vec() est appliquée telle que :

$$\mathrm{Vec}(\mathbf{X}) = \left[X_{11}, \sqrt{2}X_{12}, \ldots, \sqrt{2}X_{1d}, X_{22}, \sqrt{2}X_{23}, \ldots, X_{dd}\right], \tag{9}$$

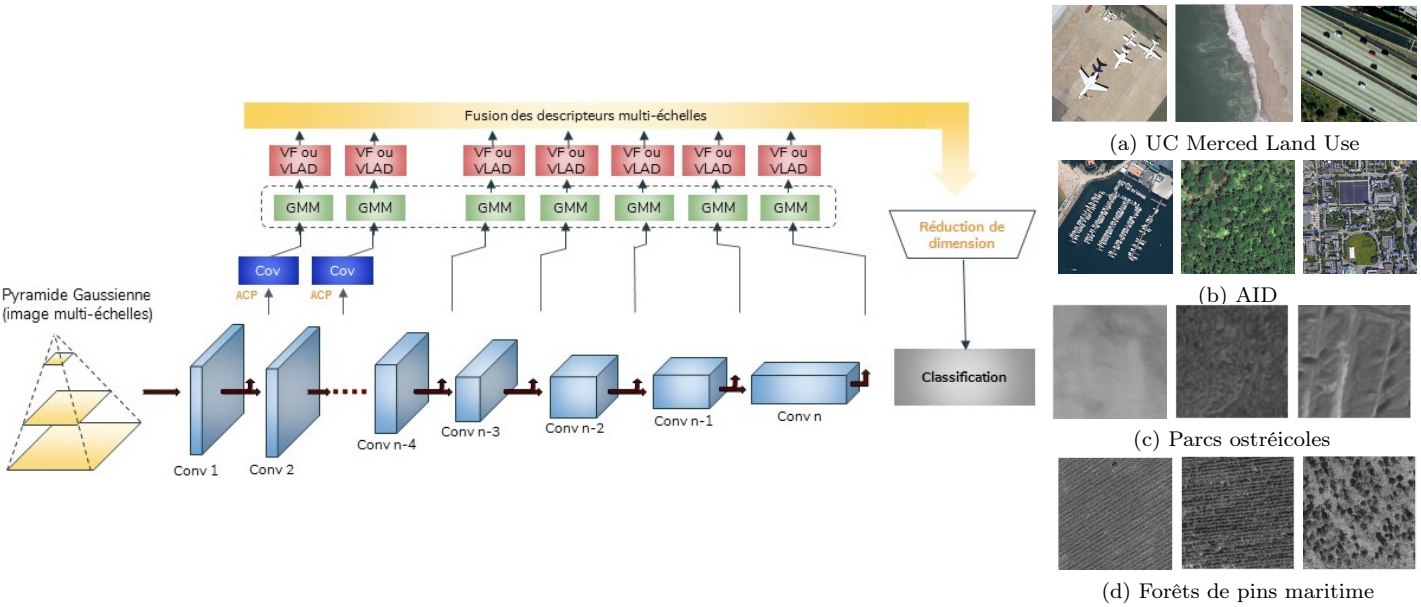
avec X_{ij} les éléments de $\mathbf{X}$ à la ligne i et la colonne j. Ces deux opérations produisent le vecteur $\mathbf{m}_n \in \mathbb{R}^{\frac{d(d+1)}{2}}$ défini par $\mathbf{m}_n = \mathrm{Vec}\left(\log_{\mathbf{M}_{ref}}(\mathbf{M}_n)\right)$.

Dans (7), $\mathbf{M}_{ref}$ représente la matrice de covariance utilisée comme référence pour définir le plan tangent. Dans [8], nous avons proposé différentes solutions pour ce point de référence comme le centre de masse ou la médiane au sens de la distance log-euclidienne. Dans la suite de ce papier, la matrice identité sera considérée comme point de référence.

3.2 Encodage LE VF des sorties des couches convolutives

Dès lors que l'ensemble des matrices de covariance sont représentées par des vecteurs $\{\mathbf{m}_n\}_{n=1:N}$, les outils classiques de la géométrie euclidienne peuvent être appliqués. En particulier, les algorithmes détaillés dans le paragraphe 2 peuvent être adaptés afin de produire les vecteurs de Fisher log-euclidiens (LE VF).

La figure 1-(i) présente l'architecture globale de l'approche hybride proposée. Comme une image en entrée



(a) UC Merced Land Use

(b) AID

(c) Parcs ostréicoles

(d) Forêts de pins maritime

FIGURE 1 – (i) Architecture hybride multi-échelles basée sur la combinaison des descripteurs VF et LE VF. (ii) Différentes bases de données utilisées pour les expérimentations

d'un CNN subit plusieurs modifications au travers les différentes couches du réseau (filtres convolutifs, étapes de pooling, ...), la taille des images en sortie des filtres convolutifs est de plus en plus réduite. Il est ainsi impossible d'extraire un nombre suffisant de matrices de covariance pour les couches profondes. Pour cela, une attention particulière est accordée au choix du réseau. Dans cette étude, le réseau CNN considéré est le réseau profond *vgg-vd-16* [15]. Ce dernier se distingue par sa capacité à préserver la dimension spatiale des images au niveau des deux premières couches convolutives. Celles-ci sont donc encodées par l'approche LE VF proposée tandis que pour les couches les plus profondes, un simple encodage par VF est considéré [13].

4 Application en télédétection

Afin d'évaluer l'efficacité de l'approche proposée en termes de performances de classification, une série d'expérimentations a été menée sur quatre bases de données différentes illustrées dans la figure 1-(ii). La base UC Merced Land Use Land Cover est constituée de 21 classes composée chacune de 100 images. Ces images de taille 256×256 pixels représentent des scènes naturelles (terrains de golf, autoroutes, plages, etc.) avec une résolution spatiale de 30 cm. Deux autres bases de données texturées sont également expérimentées. Elles sont constituées respectivement de 4 et 5 classes d'images Pléiades avec une résolution spatiale de 50 cm. La première concerne des classes d'âges de peuplement de forêts de pins maritimes tandis que la deuxième contient des images autour de parcs ostréicoles, en particulier des parcs à huîtres cultivées et des champs abandonnés. Finalement, la dernière base de données, nommée AID, est constituée de 10 000 images aériennes à grande échelle réparties dans 30 classes, de taille 600×600 pixels.

Pour les trois premières bases de données, 50 % ($p = 0.5$) des images sont utilisées pour l'apprentissage, tandis que ($p = 0.1$) pour la base AID. De plus, la dimension du dictionnaire est fixée à $K = 30$ composantes pour toutes les expérimentations [8]. En outre, pour éviter le phénomène de la fatalité de la dimension lorsque la dimensionnalité des descripteurs LE VF est élevée, une étape de réduction de dimension peut être ajoutée au processus. En sortie des couches convolutives, une analyse en composantes principales (ACP) a été introduite, cette étape permet non seulement de réduire la dimension mais aussi d'éviter la redondance des descripteurs afin de renforcer l'hypothèse de matrice de covariance diagonale dans le modèle GMM (paragraphe 2). Une deuxième étape de réduction de dimension, assurée par la méthode supervisée d'analyse discriminante à noyau (KDA), est utilisée afin de réduire la dimension des vecteurs VF et LE VF tout en préservant la séparation des classes.

Le tableau 1 résume les résultats de classification obtenus à partir de la première couche convolutive (Conv1), de la deuxième (Conv 2) ainsi que de l'ensemble des couches qui constituent l'architecture hybride multi-échelles (MS). La méthode proposée, nommée "hybride LE VF (Vgg-vd-16)" est comparée à deux méthodes de l'état de l'art. La première, "hybride VF (Vgg-vd-16)" correspond à l'approche basée sur l'encodage par VF des descripteurs du réseau CNN (voir paragraphe 2). La seconde, notée "CP (Vgg-vd-16)", est une méthode récente basée sur les statistiques d'ordre deux via le concept de *covariance pooling* [16]. Dans cette approche, les couches convolutives sont représentées par une unique matrice de covariance, tandis que dans la méthode proposée, chaque couche est modélisée par un ensemble de matrices de covariance qui sont ensuite encodées par les VF. Comme observé, les

TABLE 1 – Performance de classification pour différentes bases de données, approches hybride VF, CP et hybride LE VF

	Descripteur	Conv 1	Conv 2	MS
UC Merced $p = 0.5$	Hybride VF (Vgg-vd-16) [13]	58.2 ± 0.7 %	62.5 ± 1.2 %	**96.2 ± 0.7 %**
	CP (Vgg-vd-16) [16]	65.8 ± 1.6 %	81.9 ± 1.4 %	**96.7 ± 0.3 %**
	Hybride LE VF (Vgg-vd-16)	**70.3 ± 1.0 %**	**87.7 ± 0.7 %**	**96.7 ± 0.2 %**
Forêts de pins maritime $p = 0.5$	Hybride VF (Vgg-vd-16) [13]	85.7 ± 1.6 %	86.1 ± 1.9 %	**93.4 ± 1.5 %**
	CP (Vgg-vd-16) [16]	81.4 ± 0.3 %	89.4 ± 1.6 %	**93.9 ± 1.6 %**
	Hybride LE VF (Vgg-vd-16)	**89.9 ± 1.1 %**	**91.4 ± 0.3 %**	**95.0 ± 1.0 %**
Parcs ostréicoles $p = 0.5$	Hybride VF (Vgg-vd-16) [13]	87.7 ± 2.5 %	88.7 ± 2.1 %	**98.0 ± 1.2 %**
	CP (Vgg-vd-16) [16]	86.1 ± 1.1 %	92.3 ± 0.6 %	**98.3 ± 0.6 %**
	Hybride LE VF (Vgg-vd-16)	**94.4 ± 1.1 %**	**96.6 ± 0.7 %**	**98.4 ± 0.3 %**
AID $p = 0.1$	Hybride VF (Vgg-vd-16) [13]	33.5 ± 0.1 %	37.8 ± 0.1 %	85.8 ± 0.1 %
	CP (Vgg-vd-16) [16]	40.0 ± 0.3 %	58.8 ± 0.5 %	**87.9 ± 0.2 %**
	Hybride LE VF (Vgg-vd-16)	**52.1 ± 0.4 %**	**67.8 ± 0.4 %**	**87.6 ± 0.1 %**

quatre bases de données réagissent de manière similaire aux méthodes appliquées. Au niveau de la première et deuxième couche de convolution, un gain important se dégage dans le cas de l'approche proposée "hybride LE VF (Vgg-vd-16)" comparé aux autres méthodes. Ceci illustre l'intérêt d'introduire l'encodage des matrices de covariance afin d'élever le pouvoir discriminant de la classification. De plus, l'efficacité de la méthode peut être affirmé au vu des résultats similaires sur les quatre bases de données utilisées (pour la classification de scènes à grande échelle mais également pour la classification d'images texturées). Concernant l'architecture hybride multi-échelles (MS), les trois méthodes appliquées mènent à des résultats de classification similaires correspondants à l'état de l'art.

5 Conclusion

Ce papier a introduit une nouvelle architecture hybride basée sur l'encodage par vecteurs de Fisher (VF) des matrices de covariance extraites des sorties des couches convolutives. Pour cela, le principe des VF a d'abord été étendu au cas où les descripteurs sont des matrices de covariance permettant ainsi de prendre en compte les statistiques d'ordre 2. A cette fin, nous avons proposé d'utiliser la métrique log-euclidienne. Une application a ensuite été proposée en télédétection afin d'illustrer le potentiel de l'approche proposée pour la classification de scènes naturelles à grande échelle mais également pour des images texturées à haute résolution spatiale. En guise de perspectives, les travaux futurs concerneront la proposition d'une architecture où l'encodage par l'approche LE VF sera intégré directement dans le réseau de neurones, à l'image du réseau NetVLAD proposé dans [6]. Cela permettra à la machine d'apprendre de bout en bout à la fois les coefficients du réseau mais également les descripteurs LE VF.

6 Remerciements

Les auteurs souhaitent remercier l'équipe de la recette thématique utilisateurs du programme ORFEO (CNES) pour les images Pléiades. Les remerciements vont également au CNES, à Bordeaux Sciences Agro et au conseil régional Nouvelle Aquitaine pour le support financier.

Références

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

[2] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007, pp. 1–8.

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

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS'12 - Volume 1*. USA : Curran Associates Inc., 2012, pp. 1097–1105.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Fisher networks for large-scale image classification," in *NIPS'13*. USA : Curran Associates Inc., 2013, pp. 163–171.

[6] 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.

[7] 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.

[8] 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 Trans. Geosci. Remote Sens.*, Soumis, 2019.

[9] S. Akodad, L. Bombrun, C. Yaacoub, Y. Berthoumieu, and C. Germain, "Image classification based on log-Euclidean Fisher vectors for covariance matrix descriptors," in *(IPTA)*, Nov. 2018.

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

[11] K. Yu and M. Salzmann, "Second-order convolutional neural networks," *CoRR*, vol. abs/1703.06817, 2017.

[12] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool, "Covariance pooling for facial expression recognition," *CoRR*, vol. abs/1805.04855, 2018.

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

[14] 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.

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

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