ISPRS Journal of Photogrammetry and Remote Sensing 138 (2018) 74–85

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification

Rao Muhammad Anwer [a],*, Fahad Shahbaz Khan [b], Joost van de Weijer [c], Matthieu Molinier [d], Jorma Laaksonen [a]

[a] *Department of Computer Science, Aalto University School of Science, Finland*
[b] *Computer Vision Laboratory, Linköping University, Sweden*
[c] *Computer Vision Center, CS Dept. Universitat Autonoma de Barcelona, Spain*
[d] *VTT Technical Research Centre of Finland Ltd – Remote Sensing Team, Finland*

ARTICLE INFO

Article history:
Received 3 July 2017
Received in revised form 25 January 2018
Accepted 30 January 2018
Available online 15 February 2018

Keywords:
Remote sensing
Deep learning
Scene classification
Local Binary Patterns
Texture analysis

ABSTRACT

Designing discriminative powerful texture features robust to realistic imaging conditions is a challenging computer vision problem with many applications, including material recognition and analysis of satellite or aerial imagery. In the past, most texture description approaches were based on dense orderless statistical distribution of local features. However, most recent approaches to texture recognition and remote sensing scene classification are based on Convolutional Neural Networks (CNNs). The *de facto* practice when learning these CNN models is to use RGB patches as input with training performed on large amounts of labeled data (ImageNet). In this paper, we show that Local Binary Patterns (LBP) encoded CNN models, codenamed TEX-Nets, trained using mapped coded images with explicit LBP based texture information provide complementary information to the standard RGB deep models. Additionally, two deep architectures, namely early and late fusion, are investigated to combine the texture and color information. To the best of our knowledge, we are the first to investigate Binary Patterns encoded CNNs and different deep network fusion architectures for texture recognition and remote sensing scene classification. We perform comprehensive experiments on four texture recognition datasets and four remote sensing scene classification benchmarks: UC-Merced with 21 scene categories, WHU-RS19 with 19 scene classes, RSSCN7 with 7 categories and the recently introduced large scale aerial image dataset (AID) with 30 aerial scene types. We demonstrate that TEX-Nets provide complementary information to standard RGB deep model of the same network architecture. Our late fusion TEX-Net architecture *always* improves the overall performance compared to the standard RGB network on both recognition problems. Furthermore, our final combination leads to consistent improvement over the state-of-the-art for remote sensing scene classification.

1. Introduction

Texture analysis in real-world images, robust to variations in scale, orientation, illumination or other visual appearance, is a challenging computer vision problem with many applications, including object classification and remote sensing. Over the years, a variety of texture analysis approaches have been proposed in literature (Ojala et al., 2002; Zhang et al., 2007; Varma and Zisserman, 2010; Liu et al., 2016, 2017) to capture different properties of texture, such as spatial structure, roughness, contrast, regularity, and orientation in images. Most successful texture description methods are based on orderless distribution of local features leading to the development of several classification frameworks, including histograms of vector quantized filter responses (Leung and Malik, 1996), textons theory (Leung and Malik, 2001), bag-of-visual-words (Csurka et al., 2004) and later the Fisher Vector (Perronnin and Dance, 2007). In this paper, we tackle the issue of learning robust texture description for texture recognition *and* remote sensing scene classification.

The first problem investigated in this paper is that of texture recognition, where the task is to associate each texture image to its respective texture category. Texture recognition plays a crucial role in many applications, related to biomedical imaging, material

* Corresponding author.
E-mail address: rao.anwer@aalto.fi (R.M. Anwer).

https://doi.org/10.1016/j.isprsjprs.2018.01.023
0924-2716/© 2018 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

recognition, document image analysis, and biometrics. The problem of texture recognition can be divided into two phases: the texture description stage and the classification phase. Generally, much attention has been focused on the texture description phase since it is challenging to design powerful texture features robust to imaging conditions. One of the most successful approaches to texture description is that of Local Binary Patterns (LBP) (Ojala et al., 2002) and its variants. The standard LBP descriptor (Ojala et al., 1996) is invariant to monotonic gray scale changes and is based on the signs of differences of neighboring pixels in an image. The LBP descriptor was later extended (Ojala et al., 2002) to obtain multi-scale, rotation invariant and uniform representations and has been successfully employed in other tasks, including object detection (Zhang et al., 2011), face recognition (Ahonen et al., 2004), and remote sensing scene classification (Chen et al., 2016).

The second problem investigated in this paper is that of remote sensing scene classification. Remote sensing scene classification is a challenging and open research problem crucial for understanding high-resolution remote sensing imagery with numerous applications including vegetation mapping, urban planning, land resource management and environmental monitoring. In this problem, the task is to automatically associate a semantic class label to each high-resolution remote sensing image containing multiple land cover types and ground objects. The problem is challenging due to several factors, such as large intra-class variations, changes in illumination due to images extracted at different times and seasons, small inter-class dissimilarity and scale variations. Several existing approaches either rely on using low-level visual features (dos Santos et al., 2010; Yang and Newsam, 2013; Chen et al., 2016), such as color, shape or using combination of visual features (Luo et al., 2013; Chen et al., 2015a). Contrary to approaches based on low-level visual features, mid-level remote sensing scene classification methods tackle the problem by encoding low-level features into a holistic high-order statistical image representation. Popular mid-level approaches include bag-of-words (BOW) variants (Chen et al., 2011; Yang and Newsam, 2010), spatial extensions to BOW (Yang and Newsam, 2011; Chen and Tian, 2015a), semantic BOW using topic models (Kusumaningrum et al., 2014; Zhong et al., 2015), and unsupervised feature learning (Zhang et al., 2015; Hu et al., 2015a).

Recently, Convolutional Neural Networks (CNNs) have revolutionised computer vision, being the catalyst to significant performance gains in many vision applications, including texture recognition (Cimpoi et al., 2016) and remote sensing scene classification (Hu et al., 2015b; Penatti et al., 2015). CNNs and other "deep networks" are generally trained on large amounts of labeled training data (e.g. ImageNet (Deng et al., 2009)) with raw image pixels with a fixed size as input. Deep networks consists of several convolution and pooling operations followed by one or more fully connected (FC) layers. Several works (Azizpour et al., 2014; Oquab et al., 2014) have shown that intermediate activations of the FC layers in a deep network, pre-trained on the ImageNet dataset, are general-purpose features applicable to visual recognition tasks. Deep features based approaches have shown to provide the best results in recent evaluations for texture recognition (Liu et al., 2016) and remote sensing scene classification (Xia et al., 2017).

As mentioned above, the *de facto* practice is to train deep models on the ImageNet dataset using RGB values of the image patch as an input to the network. These pre-trained RGB deep networks are typically employed in state-of-the-art methods for texture recognition and remote sensing scene classification. Interestingly, in a recent performance evaluation for texture recognition (Liu et al., 2016), the hand-crafted LBP texture descriptor and its variants were shown to provide competitive performance compared to deep features based methods especially in the presence of rotations and several types of noise. In addition to texture recognition, LBP and its variants have been successfully employed for remote sensing scene classification (dos Santos et al., 2010; Chen et al., 2016). Moreover, the work of Levi and Hassner (2015) proposes to train CNNs on pre-processed texture coded images in addition to RGB for emotion recognition. Motivated by these observations, we investigate the impact of integrating LBP within deep learning architectures for texture recognition and remote sensing scene classification.

The combination of multiple feature streams into a single architecture has recently been a subject of intense study. It is being investigated in the context of action recognition (Simonyan and Zisserman, 2014; Cheron et al., 2015; Feichtenhofer et al., 2016), RGB-D (Hoffman et al., 2016), and multi-modal networks (Reed et al., 2016; Fukui et al., 2016). In the aforementioned multiple feature streams action recognition approaches, the spatial stream captures the appearance information by using RGB images as input to the network and the temporal stream captures the motion information by using dense optical flow images as input to the network. The spatial and motion streams are then fused since they contain complimentary information. Inspired by the success of these two-stream deep networks, we propose a two-stream deep architecture where texture coded mapped images are used as the second stream and fuse it with the normal RGB image stream. The two network streams can be fused at different stages in the deep architecture. In the first strategy, termed as late fusion, the RGB and texture streams are trained separately and combined at a later stage by fusing them at the FC layers. In the second strategy, termed as early fusion, the two streams are joined at an early stage by aggregating the RGB and texture coded image channels as an input, in order to train a joint two-stream deep model. To the best of our knowledge, we are the first to investigate these two fusion strategies, to combine RGB and texture streams, in the context of texture recognition and remote sensing scene classification.

Contributions: In this work we investigate the problem of learning robust texture description by integrating one of the most popular hand-crafted texture descriptor, Local Binary Patterns (LBP), within deep learning architectures for texture recognition and remote sensing scene classification. To this end, we propose deep models, which we call TEX-Nets, by designing a two-stream deep architecture where texture coded mapped images are used as the second stream and fuse it with the normal RGB image stream. To obtain the texture coded mapped images, we first extract LBP based codes from an image. Afterwards, as in Levi and Hassner (2015), the unordered LBP code values are mapped to points in a 3D metric space. The mapping is performed by employing Multi Dimensional Scaling (MDS) using code-to-code dissimilarity scores based on approximated Earth Mover's Distance (EMD). We further evaluate two fusion strategies, early and late fusion, to combine RGB and texture streams for texture recognition and remote sensing scene classification.

The proposed approach is first evaluated on a selection of texture benchmark datasets to demonstrate the overall effectiveness of the approach, and then applied to several remote sensing benchmark datasets to demonstrate its potential and applicability to remote sensing scene classification. The results of our experiments suggest that our late fusion TEX-Net architecture provides superior results compared to the early fusion TEX-Net architecture. Further, the proposed late fusion TEX-Net architecture *always* improves the overall performance compared to the standard RGB stream deep network architecture. Lastly, our final combination leads to performance superior to the state-of-the-art without employing finetuning or ensemble of RGB network architectures, for remote sensing scene classification.

2. Related work

Here, we briefly review the Local Binary Patterns (LBP) and its variants, deep learning and state-of-the-art in texture recognition and remote sensing scene classification.

Local binary patterns: In the field of texture recognition, local binary patterns (LBP) (Ojala et al., 2002) is one of the most commonly used texture description approaches. Besides texture recognition, LBP based texture description has been applied to other vision tasks, including face recognition (Tan and Triggs, 2007), gender recognition (Khan et al., 2014) and person detection (Wang et al., 2009). The LBP descriptor works by thresholding intensity values of a pixel around its neighborhood. The threshold is computed from the intensity of each neighborhood's center pixel. A circular symmetric neighborhood is employed by interpolating the locations not exactly at the center of a pixel. A variety of LBP variants have been proposed in literature, including Local Ternary Patterns (Tan and Triggs, 2010), Local Binary Pattern Variance (Guo et al., 2010a), Noise Tolerant Local Binary Patterns (Fathi and Nilchi, 2012), Completed Local Binary Patterns (Guo et al., 2010b), Extended Local Binary Patterns (Liu et al., 2012) and Rotation Invariant Local Phase Quantization (Ojansivu et al., 2009). In addition to the introduction of different LBP variants, the fusion of LBP descriptor with color features have also been investigated in previous studies (Maenpaa and Pietikainen, 2004; Khan et al., 2015).

Deep learning: In recent years, Convolutional Neural Networks (CNNs) (LeCun et al., 1989) have shown to provide excellent performance for many computer vision tasks. CNNs are generally trained using large amount of labeled training samples and take fixed sized RGB images as input to a series of convolution, normalization and pooling operations (termed as layers). The network typically ends with several fully-connected (FC) layers, used to extract features for recognition. Several attempts have been made to improve deep network architectures, including increasing the depth of the network by introducing additional convolutional layers (Simonyan and Zisserman, 2015; He et al., 2016). In addition to RGB based appearance networks, other modalities such as motion and depth have also been used to construct multi-cue deep networks for action recognition (Simonyan and Zisserman, 2014) and RGB-D object recognition (Eitel et al., 2015).

Deep learning for remote sensing image analysis: In recent years, deep learning methods have made a breakthrough for satellite image analysis, with several works published in the major remote sensing journals. The most notable applications of deep neural networks (DNNs) in remote sensing include land cover classification with optical images (Chen et al., 2014a; Romero et al., 2016; Matthieu Molinier and Laaksonen, 2007), hyperspectral image analysis (Chen et al., 2014b, 2015b; Tuia et al., 2015) or Synthetic Aperture Radar (SAR) image analysis (Geng et al., 2015).

A large majority of published works use DNNs trained on patches extracted from satellite images. DNNs are usually not trained on databases of full sized satellite images (1 to several GB per image) due to memory limitations, even on powerful GPU servers. CNNs are the most commonly used deep learning architectures for the classification of optical (Chen et al., 2014a) and SAR (Geng et al., 2015) satellite images. Because large datasets of satellite images with high quality labels are not easily available, most of the earlier works utilized pre-trained DNNs that were trained on computer vision benchmark datasets (ImageNet), not on satellite images (Marmanis et al., 2016).

Texture recognition: A variety of texture recognition approaches have been proposed in literature (Liu et al., 2016; Pietikainen and Zhao, 2015). The work of Varma and Zisserman (2010) proposes a statistical approach to model textures based on the joint probability distribution of filter responses. The work of Chen et al. (2010) proposes an approach based on Weber's law which consists of two components: differential excitation and orientation. An image is represented by the concatenation of these two components in a single representation. The work of ul Hussain and Triggs (2012) introduces an approach that uses lookup-table based vector quantization for texture description. A set of low and mid-level perceptually inspired visual features are introduced by Sharan et al. (2013) for texture recognition. A multi-resolution framework based on LBP is proposed by Ojala et al. (2002) for rotation invariant texture recognition. As discussed earlier, LBP is one of the most successful approaches for texture recognition with several variants existing in literature (Guo et al., 2010c; Ylioinas et al., 2013, 2012).

Other than LBP and its variants, bag-of-words based representations employing SIFT features and Fisher Vector encoding scheme have shown promising results for texture recognition (Cimpoi et al., 2014). Recently, deep features have also been investigated for texture recognition. Bruna and Mallat (2013) introduce the wavelet convolutional scattering network (ScatNet), where no learning is required and convolutional filters are defined as wavelets. The work of Chan et al. (2014) proposes a deep network based on multistage principal component analysis (PCANet). The work of Cimpoi et al. (2016) proposes to use the convolutional layers of the deep networks as dense local descriptors encoded with Fisher Vector to obtain the final image representation.

Our approach: As discussed above, most existing hand-crafted approaches employ LBP and its variants for texture description. On the other hand, deep learning based approaches have shown promising results for texture recognition and remote sensing scene classification. Despite the success of deep features, the hand-crafted LBP texture descriptor and its variants have been shown to provide competitive performance compared to deep feature based methods especially in the presence of rotations and several types of noises in a recent performance evaluation for texture recognition (Liu et al., 2016). Moreover, the deep features based texture recognition and remote sensing scene classification methods employ deep networks pre-trained on the ImageNet dataset using *RGB images* as input. This motivates us to investigate the impact of integrating texture features, in particular, the popular hand-crafted LBP texture descriptor within deep learning architectures. We investigate fusion strategies by constructing a two-stream deep architecture where texture coded mapped images are used as the second stream and fuse it with the normal RGB image stream. To the best of our knowledge, we are the first to investigate the two fusion strategies in a two-stream deep architecture, to combine RGB and texture streams, in the context of texture recognition and remote sensing scene classification. This paper is an extended version of our earlier work (Anwer et al., 2017). We have extended our experiments by evaluating the proposed approach for remote sensing scene classification application with results on four challenging benchmarks. In addition, we also provide an analysis of our two-stream deep architecture on the ImageNet dataset.

3. Binary patterns encoded convolutional neural networks

Here, we first describe the construction of deep models based on texture coded mapped images. Afterwards, we investigate different strategies to fuse the texture coded mapped stream with the normal RGB image stream.

3.1. Mapped LBP codes

As discussed earlier, Local Binary Patterns (LBP) has shown competitive performance for texture recognition and is one of the most commonly employed approaches for texture description. LBP features describe the neighborhood of a pixel by its binary derivatives. These binary derivatives are then used to form a short code to describe the neighborhood of the pixel. The short LBP codes are binary numbers (lower than threshold (0) or higher than the threshold (1)). Each LBP code can be considered as a micro-texton since each pixel is assigned a code of the texture primitive with its best local neighborhood match. Several local primitives are detected by the LBP operator, including flat areas, edges, corners, curves, and edge ends. The primitive version of LBP operator considered only the eight-neighbors of a pixel, while using the center pixel value as a threshold. Later variants extended the primitive LBP operator to consider all circular neighborhoods with any number of pixels. Given an image $f(a_c, b_c)$ of size $H \times W$, with $(a_c \in \{0, \ldots, H-1\}, b_c \in \{0, \ldots, W-1\})$. Here, (a_c, b_c) are the coordinates of the center pixel of a circular local neighborhood (P, R), where P denotes the number of sampling points and $R(R > 0)$ is the circle radius of the of local neighborhood. The LBP code (a P-bit word) describing the local image texture around the center pixel is computed as,

$$LBPC_{P,R}(a_c, b_c) = \sum_{p=0}^{P-1} s(f(a_p, b_p) - f(a_c, b_c))2^p, \tag{1}$$

where the thresholding function $s(t)$ is defined as:

$$s(t) = \begin{cases} 0 & \text{for } t < 0 \\ 1 & \text{for } t \geqslant 0. \end{cases} \tag{2}$$

The standard LBP computation results in 2^p distinct values for the LBP code. In case of an 8 pixel neighborhood, the LBP code computation results in a binary string of eight-bit numbers between 0 and 255. The final image representation is obtained by computing the histogram as a distribution of LBP codes over an entire image region. The resulting feature vector normalizes for translation and is invariant to monotonic changes in the gray scale.

As discussed above, the LBP codes are generally pooled as histogram representations and employed as an input to a discriminative classifier, such as Support Vector Machines (SVMs). Instead, due to the overwhelming recent success of deep learning, it is worth investigating to integrate the strength of LBP descriptor within the CNN architectures. A straightforward integration strategy is to train deep models by directly using LBP codes as CNN inputs. However, such a strategy is not applicable since the convolution operations, equivalent to a weighted average of the input values, performed within CNN models are unsuitable for the unordered nature of the LBP code values.

Recently, the work of Levi and Hassner (2015) provides a solution to this problem within the context of texture description for emotion recognition. They propose to map the LBP codes to points in a 3D metric space in which the Euclidian distance approximates the distance between the LBP codes. After the transformation of the LBP codes they can be averaged together using convolution operations within CNN models.

The method is based on defining a distance $\delta_{j,k}$ between the LBP codes $LBPC_j$ and $LBPC_k$. The authors of Levi and Hassner (2015) choose the Earth Movers Distance (EMD) (Rubner et al., 2000) because it accounts for both the different bit values and their locations. Having defined the distance between the LBP codes, it is now possible to look for a mapping of the LBP codes into a D-dimensional space which approximately preserves this distance. This mapping can be found by applying Multi Dimensional Scaling (MDS) (Borg and Groenen, 2005), such that:

$$\delta_{j,k} \approx \|L_j - L_k\| = \|MDS(LBPC_j) - MDS(LBPC_k)\|. \tag{3}$$

where $L_j = MDS(LBPC_j)$ is the mapping of code j into the D-dimensional space. Applying this mapping allows us to transfer the LBP codes into a representation which can be used as input to a CNN. In Levi and Hassner (2015) they experimented with the optimal dimensionality D and found that good results were obtained with $D = 3$. In this work, we use the same settings and in addition also investigate an early fusion scheme with $D = 1$. We refer to Levi and Hassner (2015) for more details. Fig. 1 shows an example image converted to LBP codes (middle). The LBP codes are mapped to a 3D metric space (right) and normalized before used as an input to CNNs.

3.2. Texture coded two-stream deep architecture

As described earlier, the de facto standard when training deep models is to use raw RGB pixels values of an image as input. These RGB based deep networks have achieved state-of-the-art results for texture recognition recently (Cimpoi et al., 2016) and remote sensing scene classification (Hu et al., 2015b; Penatti et al., 2015). In this work, we investigate to what extent texture coded deep networks complement the standard RGB based CNN models in two classification problems: texture recognition and remote sensing scene classification. To this end, we design a two-stream deep architecture, referred as TEX-Nets, using both texture coded mapped images (Section 3.1) and raw RGB pixel values. Our TEX-Nets models are trained on the ImageNet ILSVRC-2012 dataset (Deng et al., 2009). We employ two different architectures to validate our approach: the VGG-M architecture (Chatfield et al., 2014) which is similar to Zeiler and Fergus network (Zeiler and Fergus, 2014) and the ResNet architecture (He et al., 2015). The VGG-M network comprises of five convolutional and three fully-connected (FC) layers. The VGG-M network takes as input an image of 224 × 224 pixels. The first convolutional layer employs smaller stride (1) and receptive field (or the filter size). The second convolutional layer uses a relatively larger stride (2 compared to 1). The number of convolution filters is 96 in the first convolutional layer, 256 in the second convolutional layer and 512 in the third and last convolutional layers. During training, the learning rate is set to 0.001, a weight decay that acts as a regularizer and helps reducing

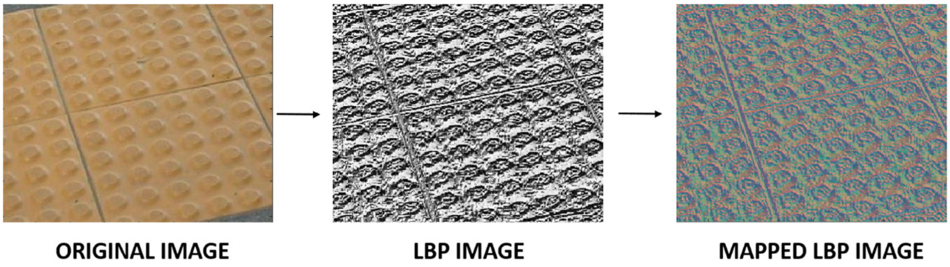

Fig. 1. Example of the texture coded mapped image (visualized here in color). The mapped LBP image is obtained by converting LBP codes (shown as grayscale values) into a 3D metric space.

the training error of the model is set to 0.0005. The momentum rate is associated with the gradient descent algorithm used to minimize the objective function and is set to 0.9. We also employ the ResNet-50 architecture (He et al., 2015) which is a 50 layer Residual Network. This architecture is based on residual learning framework that facilitates efficient training of deeper networks by reformulating the layers as learning residual functions with reference to the layer inputs. The ResNet-50 architecture takes as input an image of 224 × 224 pixels. For the first 30 training iterations, the learning rate is set to 0.1. For the second and the last 30 training iterations, the learning rate is set to 0.01 and 0.001 respectively. The momentum and the weight decay is set to 0.9 and 0.0001 respectively.

Next, we investigate strategies to fuse the two network streams at different stages in the deep architectures.

Late fusion: In this strategy, both standard (RGB) and texture coded network streams are trained separately on the ImageNet dataset. The standard RGB network stream takes RGB values as input, whereas the second network stream takes texture coded mapped images as input. These texture coded mapped images are obtained by first employing the LBP encoding that converts intensity values in an image to one of the 256 LBP code values. The LBP code values are then mapped into a 3D metric space (Section 3.1). The resulting 3-channel texture coded mapped images are then used as input to CNN models. Despite being efficient to compute, the texture coded mapped images still introduce a bottleneck if done on-the-fly. We therefore pre-compute these texture coded mapped images before training the deep network. Once separately trained, the RGB and texture coded network streams are combined at a later stage by fusing them in the FC layers in the VGG-M architecture. In case of ResNet architecture, late fusion is performed before the softmax loss. The two-stream late fusion strategy has been previously used in action recognition to combine spatial (RGB) and temporal (flow) information (Simonyan and Zisserman, 2014; Cheron et al., 2015).

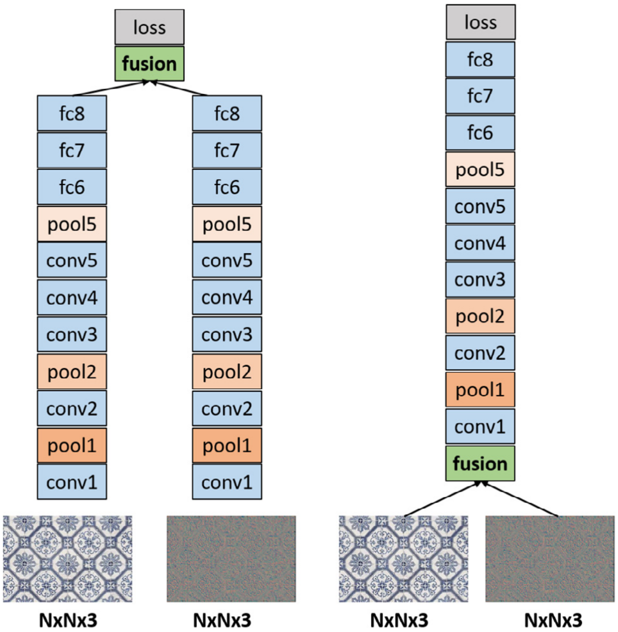

Fig. 2. Two-stream deep fusion VGG-M architectures. The left example shows late fusion architecture where the deep models trained using RGB and texture coded mapped images are kept separately. The point of fusion, to combine the two network towers, is in the FC layer. The right example shows early fusion architecture where the point of fusion is the input to the network. As a result, a joint deep model is trained by aggregating the RGB and texture image channels as an input to the network.

Early fusion: Other than late fusion, we also investigate an alternative strategy, termed as early fusion, where the point of fusion is the input to the network. In the early fusion based two-stream network architecture, a joint deep model is trained by aggregating the RGB and texture coded mapped image channels as an input to the deep network. As a result, the input to CNN is an image of 224 × 224 × 6 dimensions. We employ same early fusion strategy for both VGG-M and ResNet architectures. We also investigated converting the 3-channel mapped coded images into a single channel and combining it with the three RGB channels. In both networks, the filters are learned jointly on the RGB and texture coded images. Fig. 2 shows both early and late fusion based two-stream deep fusion VGG-M architectures designed to combine the color and texture image streams.

Training TEX-Nets on ImageNet: As described earlier, we train our TEX-Nets from scratch on the ImageNet 2012 dataset employed in ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The dataset consists of 1000 object classes and 1.2 million training images, 50,000 validation images, and 100,000 test images. On this dataset, the results are measured by top-1 and top-5 error rates. The error rates are computed from the predictions using the deep network and obtaining the predicted class multinomial distributions. The top-5 error is the fraction of test images for which the true label is not among the five labels (the 5 predictions with the highest probabilities) considered most probable by the deep model. The top-1 error is computed by evaluating if the top class (the one having the highest confidence) is the same as the correct (target) label. Table 1 shows the classification performance comparison, based on VGG-M architecture, of our early and late fusion based two-stream deep TEX-Net architectures with the standard RGB deep network on the ILSVRC 2012 dataset. The standard baseline RGB network achieves top-1 and top-5 errors of 37.6% and 15.9% respectively. Our late fusion deep architecture significantly reduces the error with an absolute reduction of 3.2% in the top-5 error, compared to the standard RGB network. The late fusion architecture results in increasing the number of network parameters by a factor of 1.4, compared to the standard RGB. We therefore also train a six channel early fusion network by increasing the network depth with a factor of 1.4, resulting in same number of parameters as late fusion. This improves the results for the six channel early fusion architecture. However, it still provides inferior results (35.3 top-1 error) compared to the late fusion architecture (34.4 top-1 error).

We further validate the effectiveness of our late fusion two-stream approach by employing the ResNet-50 architecture. Table 1 shows the classification performance comparison of our approach and the standard RGB network. Both networks are trained from scratch on the ImageNet dataset. The standard baseline RGB network achieves top-1 and top-5 errors of 25.4% and 8.0% respectively. Our late fusion based TEX-Net ResNet architecture (TEX-Net-LF) reduces the error with top-1 and top-5 errors of 23.7% and 7.0% respectively.

Fig. 3 shows 20 object categories from the ImageNet dataset where our late fusion two-stream deep architecture provides the largest reduction in the top-5 error rate, compared to the standard RGB deep network. For majority of the depicted classes it is likely that a good texture representation is crucial for correct classification. Consequently, the aforementioned results suggest that our late fusion two-stream deep architecture provides superior results compared to both standard RGB and early fusion.

4. Experimental results

Here, we start by evaluating our TEX-Net deep models for the texture recognition problem. We then provide a comparison of

Table 1
Classification performance comparison of our two-stream deep TEX-Net architectures with the standard RGB network on the ImageNet ILSVRC 2012 validation data. In case of VGG-M architecture, we show comparison with both early and late fusion TEX-Net models: early fusion architecture aggregating the RGB and 3 mapped coded channels (TEX-Net-EF-6ch), early fusion architecture aggregating the RGB and a single mapped coded channel (TEX-Net-EF-4ch) and the late fusion architecture (TEX-Net-LF) combining separate streams of RGB and texture networks. We also show results based on only mapped coded images (TEX-Net Standard), without color information. In case of ResNet architecture, we show the comparison between our late fusion approach and the standard RGB network.

Method	Architecture	Channels	Top-1 error (%)	Top-5 error (%)
Standard RGB (baseline)	VGG-M	3	37.6	15.9
TEX-Net standard	VGG-M	3	45.8	21.9
TEX-Net-EF-6ch	VGG-M	6	39.3	17.7
TEX-Net-EF-4ch	VGG-M	4	37.1	15.5
TEX-Net-LF	VGG-M	6	34.4	13.8
Standard RGB (baseline)	ResNet	3	25.4	8.0
TEX-Net-LF	ResNet	6	**23.7**	**7.0**

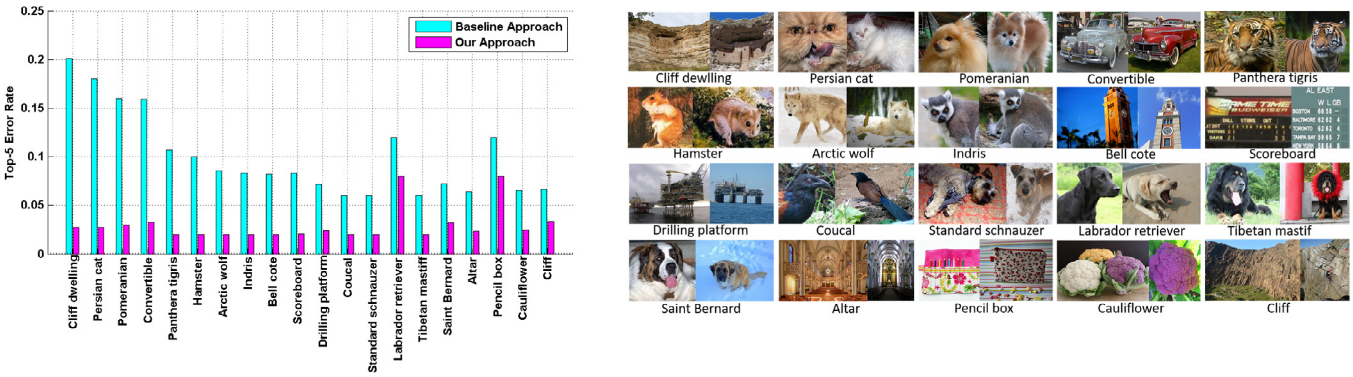

Fig. 3. Object categories in the ImageNet dataset where our late fusion two-stream deep architecture provides significant reduction in the top-5 error compared to the baseline standard RGB deep network. On the left, we show the comparison (in top-5 error) and on the right, we show example images from these object categories (left to right). Both approaches are based on VGG-M architecture.

our approach with the standard RGB based deep network in the remote sensing scene classification task. Finally, we compare the performance of our approach with state-of-the-art remote sensing scene classification results reported in literature.

4.1. Texture recognition

We evaluate our approach by performing experiments on four challenging texture datasets: DTD, KTH-TIPS-2a, KTH-TIPS-2b and Texture-10. Fig. 4 shows example images from the four texture datasets.

DTD: The DTD dataset consists of 5640 images from 47 texture classes, collected from the web. Each texture class consists of 120 images with the dataset equally divided into training, validation and test. The training and test splits are provided by the authors.

KTH-TIPS-2a: The KTH-TIPS-2a dataset consists of 11 texture classes. The 4752 images are captured at 9 different scales, 3 poses and 4 different illumination conditions. Similar to previous works (Chen et al., 2010; Caputo et al., 2005; Sharma et al., 2012), average classification performance is reported over the 4 test runs. In each run, images from 1 sample are used for testing while the images from the remaining 3 samples are used as a training set.

KTH-TIPS-2b: The KTH-TIPS-2b dataset consists of 11 texture categories. Here, images from 1 sample are used for training while all the images from remaining 3 samples are used for testing in each test run.

Texture-10: The Texture-10 dataset consists of 400 images of 10 different texture categories. For each texture category, 25 images are used for training and 15 images are used for testing.

Experimental setup: As discussed earlier, both the TEX-Net networks and the standard RGB deep network are trained from scratch on ImageNet 2012 training set. The deep models are trained by employing the Matconvnet library (Vedaldi and Lenc, 2015). We evaluate our VGG-M architecture based deep models, pre-trained on ImageNet, as feature extractors on texture datasets. We therefore remove the last fully-connected layer (FC8) of the VGG-M networks which performs 1000-way ImageNet (ILSVRC) classification, and instead use 4096 dimensional activations from the FC7 (second last) layer as image features. The resulting image features are L_2-normalised and input to a linear SVM classifier. Throughout our experiments, we fixed the weights (no fine-tuning) of all the pre-trained deep VGG-M networks for fair comparison. In all cases (datasets), the results are reported as average recognition accuracy over all texture categories in a texture dataset. The classification is performed by employing one-versus-all SVMs with linear kernel. The category label from the classifier providing the highest confidence is assigned to the test instance. The overall classification results are then obtained by calculating the average of the classification scores of all texture classes in a dataset.

In case of ResNet architecture, we fine-tuned both the standard RGB and our late fusion two-stream model to perform classification in an end-to-end fashion. For fine-tuning on each dataset, we use the training samples with a batch size of 80 and a momentum value of 0.9. The learning rate is set to 0.005.

4.2. Baseline comparison

We compare our TEX-Net deep models with the standard RGB based CNN approach to validate whether RGB and texture coded mapped images contain complementary information. We further evaluate both early and late fusion two-stream deep architectures (Section 3.2) for combining texture and color information. For fair comparison, we use the same network architecture (VGG-M)

Fig. 4. Example images from the four texture datasets: DTD, KTH-TIPS-2a, KTH-TIPS-2b and Texture-10.

Table 2
Comparison (in %) of our approaches with the standard RGB deep network on four texture datasets. In case of VGG-M architecture, we show comparison with our different TEX-Net models: based on only mapped coded images (TEX-Net Standard), early fusion two-stream architectures combining either the RGB and 3 mapped coded channels (TEX-Net-EF-6ch) or RGB and a single mapped coded channel (TEX-Net-EF-6ch), and the late fusion architecture (TEX-Net-LF) combining standard RGB and TEX-Net standard networks. In case of ResNet architecture, we show the comparison between our late fusion approach and the standard RGB network. For both VGG-M and ResNet architectures, our late fusion approach *always* outperforms the corresponding baseline standard RGB network.

	Architecture	DTD	KTH-TIPS-2a	KTH-TIPS-2b	Texture-10
Standard RGB	VGG-M	63.4 ± 0.7	81.8 ± 5.1	72.9 ± 2.1	87.3
TEX-Net Standard	VGG-M	55.9 ± 1.1	68.6 ± 5.3	60.2 ± 2.9	81.7
TEX-Net-EF-6ch	VGG-M	64.0 ± 0.8	82.6 ± 5.5	73.6 ± 2.6	89.1
TEX-Net-EF-4ch	VGG-M	64.6 ± 0.9	83.4 ± 5.3	73.8 ± 2.7	89.3
TEX-Net-LF	VGG-M	68.2 ± 0.8	85.3 ± 5.6	75.5 ± 2.7	91.3
Standard RGB	ResNet	69.6 ± 0.7	83.3 ± 5.1	75.2 ± 2.9	90.1
TEX-Net-LF	ResNet	**73.6** ± **0.6**	**88.3** ± **5.3**	**78.0** ± **2.8**	**92.3**

together with the same set of parameters for all the deep models. Table 2 shows the baseline comparison on four texture datasets. In case of VGG-M architecture, the standard RGB deep network provides a mean accuracy of 63.4% on the DTD dataset. The two early fusion based two-stream deep architectures (TEX-Net-EF-6ch and TEX-Net-EF-4ch) slightly improve the accuracy over the standard RGB, with mean classification scores of 64.0% and 64.6% respectively. The image representation based on the TEX-Net standard model provides a classification score of 55.9%. On this dataset, the best results are obtained with our late fusion based two-stream deep ResNet architecture. On the KTH-TIPS-2a dataset, the standard RGB deep network provides a mean classification rate of 81.8%. Our TEX-Net standard model based on texture coded mapped images provides a classification score of 68.6%. The two early fusion based two-stream deep architectures (TEX-Net-EF-6ch and TEX-Net-EF-4ch) provide slight improvement in performance over standard RGB, with mean recognition scores of 82.6% and 83.4% respectively. Our late fusion based two-stream deep architecture achieves a mean classification rate of 85.3%. When using the ResNet architecture, our late fusion approach provides superior results compared to the standard RGB network.

In case of VGG-M architecture, the baseline RGB deep network provides a mean accuracy of 72.9% on the KTH-TIPS-2b dataset. The two early fusion based two-stream deep architectures (TEX-Net-EF-6ch and TEX-Net-EF-4ch) achieve mean classification scores of 73.6% and 73.8% respectively. When using the ResNet architecture, our late fusion based deep network provides a gain of 2.8% over the standard RGB network. Finally, on the Texture-10 dataset, the standard RGB deep network achieves a mean classification score of 87.3% with VGG-M architecture. Our late fusion based two-stream deep VGG-M architecture obtains a mean accuracy of 91.3%, leading to a gain of 4.0% compared to the standard RGB VGG-M network. The best results are obtained using our late fusion approach with ResNet architecture. Fig. 5 shows a VGG-M architecture based visualization of filter weights (on the left) from the RGB and TEX-Net model respectively and a visualization of activations (on the right) with the highest energy from the conv3 layer of the RGB (top row) and TEX-Net (bottom row) networks on an example texture image. In conclusion, the results suggest a robust description of texture features with the proposed approach, which we then apply to remote sensing benchmark datasets.

4.3. Remote sensing scene classification

We evaluate our approach by performing experiments on four challenging remote sensing scene classification datasets: UC-Merced, WHU-RS19, RSSCN7 and the recently introduced AID.

UC-Merced is a publicly available dataset (Yang and Newsam, 2010) consisting of 2100 aerial scene images with pixel resolution of one foot, downloaded from the United States Geological Survey (USGS) National Map. The images were downloaded from 20 regions across the USA: Buffalo, Boston, Birmingham, Columbus, Dallas, Houston, Harrisburg, Jacksonville, Las Vegas, Los Angeles, Miami, New York, Napa, Reno, San Diego, Santa Barbara, Seattle, Tampa, Tucson, and Ventura. The images in the dataset are cropped into 256 × 256 pixels, equally divided into 21 classes: agriculture, airplane, baseball diamond, beach, buildings, chaparral (shrubland/heathland), dense residential, forest, freeway, golf course, harbor,

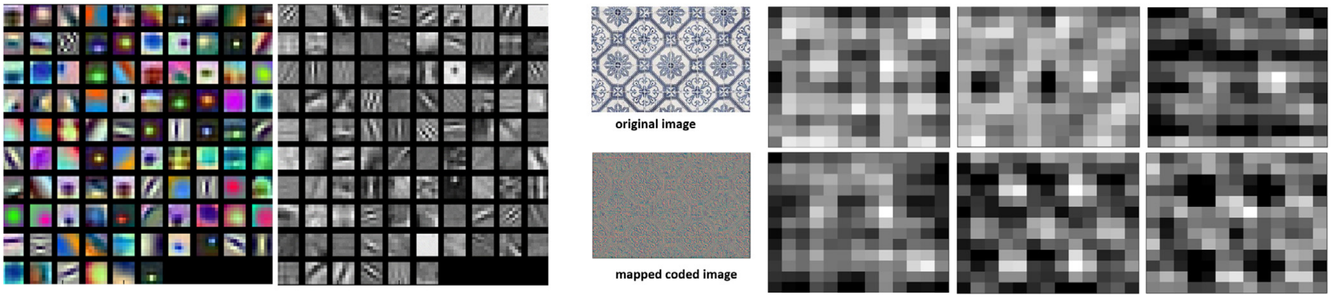

Fig. 5. On the left, visualization of filter weights from the RGB and TEX-Net VGG-M model with mapped coded texture information respectively. On the right, visualization of activations with highest energy from the conv3 layer of RGB (top row) and TEX-Net (bottom row) networks on an example texture image. The TEX-Net model is trained on the texture coded mapped images (visualized here in color), obtained by converting LBP codes into a 3D metric space. In both cases, the models are based on VGG-M architecture.

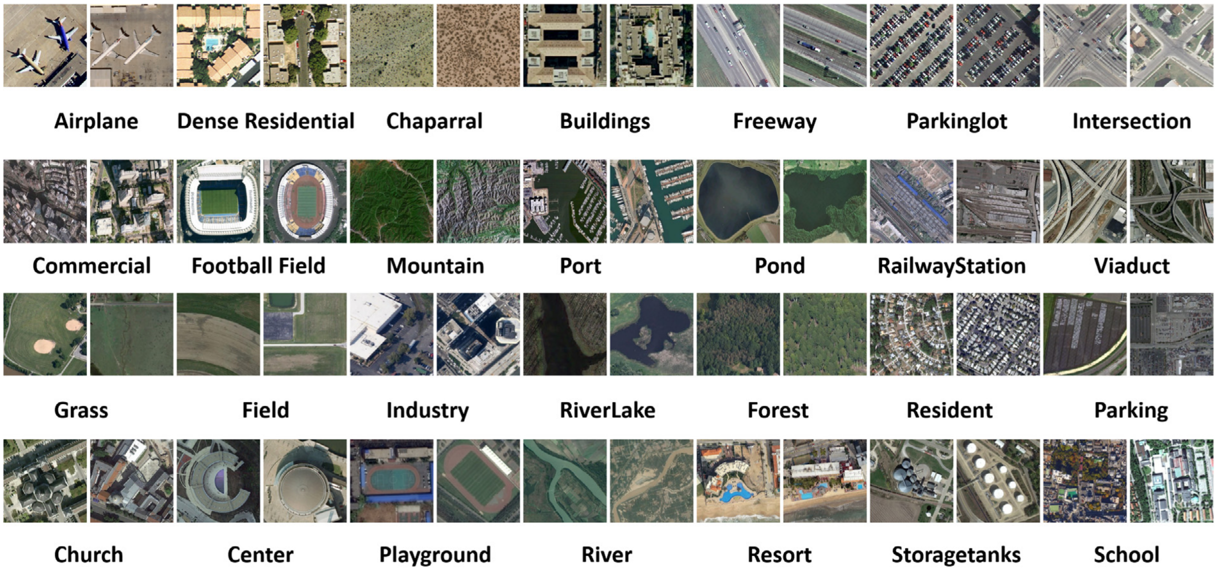

Fig. 6. Example images from the four remote sensing scene classification datasets from top to bottom: UC-Merced, WHU-RS19, RSSCN7 and the recently introduced AID.

intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. The dataset is challenging with a variety of spatial land-use patterns with a significant overlap among several categories, such as medium residential, sparse residential and dense residential. These overlapping categories only differ in the density of structures.

WHU-RS19 is a publicly available dataset (Sheng et al., 2012) consisting of 950 high spatial resolution aerial images collected from Google Earth imagery. The images in the dataset are of size 600 × 600 pixels, 50 samples per category and equally divided into 19 scene classes: airport, beach, bridge, river, forest, meadow, pond, parking, port, viaduct, residential area, industrial area, commercial area, desert, farmland, football field, mountain, park, and railway station. The dataset is challenging since images within each scene class are collected from different regions around the world with scale variations and different lighting conditions.

RSSCN7 is a publicly available dataset (Zou et al., 2015), released in 2015, consisting of 2800 aerial scene images. The images are divided into 7 scene classes: grassland forest, farmland, parking lot, residential region, industrial region, river, and lake. Each scene class comprises of 400 images, where each image has a size of 400 × 400 pixels. The dataset is challenging since images in each category are sampled at four different scales with different imaging angles.

AID is a recently introduced publicly available large-scale aerial image dataset (Xia et al., 2017). The dataset consists of 10,000 images and 30 aerial scene categories: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks and viaduct. Unlike other aerial scene datasets, such as the UC-Merced dataset, the images in the AID dataset are collected from Google Earth imagery using different remote imaging sensors. The dataset is challenging since images in each scene category are collected from different countries around the world including China, USA, UK, France, Italy, and Germany. Further, the images are collected under varying imaging conditions (time and seasons), thereby further complicating the task of aerial scene classification. Fig. 6 shows example images from the four remote sensing scene classification datasets.

Experimental setup: We follow the standard protocol (Xia et al., 2017) to evaluate our approach on benchmark datasets. The performance is measured in terms of mean classification accuracy over all scene categories in a dataset. The classification accuracy is computed as $\frac{S_p}{S_t}$, where S_p is the number of correct predictions (images) in the test set and S_t is the total number of samples (images) in the test set. To compute the accuracy, each dataset is randomly split into training and test sets for evaluation. The evaluation procedure is then repeated ten times for a reliable performance comparison. The final results are reported as the mean and standard deviation over the ten runs. Following Xia et al. (2017), in case of UC-Merced dataset, the ratio of training

Table 3
Baseline comparison of our Tex-Net models (overall accuracy (OA) in %) with the standard RGB network on UC-Merced, WHU-RS19, RSSCN7 and AID datasets. Our late fusion based two-stream deep ResNet architecture *always* outperforms the standard baseline RGB deep ResNet.

Method	Architecture	UC-Merced (50%)	UC-Merced (80%)	WHU-RS19 (40%)	WHU-RS19 (60%)	RSSCN7 (20%)	RSSCN7 (50%)	AID (20%)	AID (50%)
Standard RGB	VGG-M	94.13 ± 0.38	95.40 ± 0.91	96.01 ± 0.54	96.57 ± 0.87	86.0 ± 0.63	88.8 ± 0.55	87.70 ± 0.33	90.29 ± 0.37
TEX-Net Standard	VGG-M	91.25 ± 0.58	92.91 ± 0.88	92.41 ± 0.76	94.53 ± 0.77	83.64 ± 0.68	86.30 ± 0.75	82.0 ± 0.23	85.25 ± 0.45
TEX-Net-EF-6ch	VGG-M	94.36 ± 0.90	95.27 ± 0.96	94.71 ± 0.77	96.0 ± 0.74	85.65 ± 0.79	88.70 ± 0.78	86.84 ± 0.34	89.68 ± 0.19
TEX-Net-EF-4ch	VGG-M	94.22 ± 0.5	95.31 ± 0.69	95.78 ± 0.87	96.40 ± 0.81	86.77 ± 0.76	89.61 ± 0.54	87.32 ± 0.37	90.0 ± 0.33
TEX-Net-LF	VGG-M	95.89 ± 0.37	96.62 ± 0.49	97.61 ± 0.36	98.0 ± 0.52	88.61 ± 0.46	91.25 ± 0.58	90.87 ± 0.11	92.96 ± 0.18
Standard RGB	ResNet	96.22 ± 0.38	96.80 ± 0.51	97.83 ± 0.38	98.24 ± 0.53	90.23 ± 0.43	93.12 ± 0.55	92.33 ± 0.13	94.91 ± 0.19
TEX-Net-LF	ResNet	**96.91 ± 0.36**	**97.72 ± 0.54**	**98.48 ± 0.37**	**98.88 ± 0.49**	**92.45 ± 0.45**	**94.0 ± 0.57**	**93.81 ± 0.12**	**95.73 ± 0.16**

to test images was set to 50:50 and 80:20 respectively, with the images randomly selected for each category. In the case of WHU-RS19, the ratio of training to test samples was set to 40:60 and 60:40 respectively. In the case of RSSCN7 and AID datasets, the ratio of the training set was fixed to 20% and 50% per class respectively. As in texture recognition (Section 4.1), we use 4096-dimensional activations from the FC7 (second last) layer as image features, where the resulting image features are L_2-normalised and input to a linear SVM classifier. Consequently, we fine-tuned both our late fusion based approach and the standard RGB ResNet architecture to perform end-to-end remote sensing scene classification. For fine-tuning ResNet models, we used the same parameter settings as in texture recognition experiments.

4.4. Baseline comparison

Table 3 shows the baseline comparison on four remote sensing scene classification datasets. In case of VGG-M architecture, the two early fusion based two-stream architectures provide slightly inferior performance compared to the baseline RGB network. As in texture recognition, the best results are obtained when using our late fusion based two-stream deep architecture approach, providing consistent improvements over the baseline standard RGB deep network for both VGG-M and ResNet architectures. A large gain in classification accuracy is achieved on the RSSCN7 and the large scale AID datasets. The RSSCN7 dataset comprises of several natural scene categories, such as grassland forest and farmland where texture features provide valuable complementary information to color features when other spectral channels besides RGB (like Near-Infrared) are not available. Similarly, the recently introduced large scale AID dataset consists of both natural scene types (farmland and forest) and man made scene categories (medium residential, sparse residential and school). Our late fusion approach achieves favorable results compared to the baseline RGB deep network. Fig. 7 shows per-class classification performance comparison of our late fusion approach compared to the baseline RGB deep network, when using the VGG-M architecture. Our approach provides consistent improvement in performance on most scene categories.

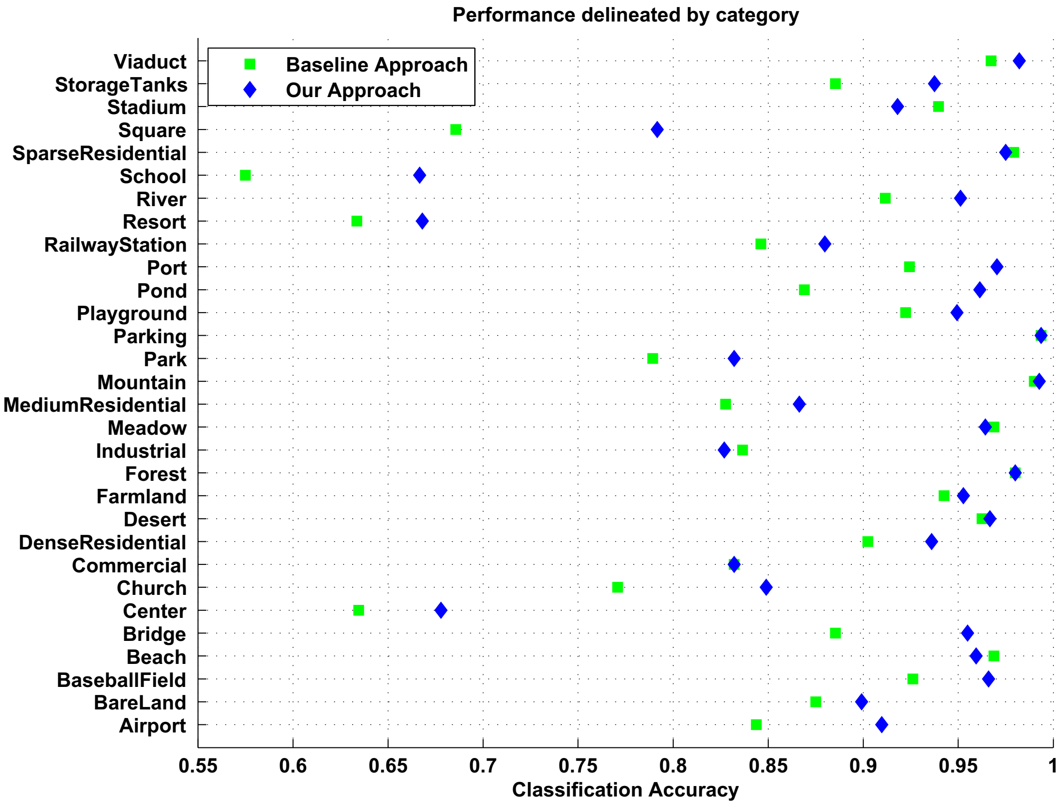

Fig. 7. Per-category performance comparison of our approach compared to the baseline RGB deep network on the AID dataset. Both the networks are based on the VGG-M architecture. Our approach improves the classification performance on most scene categories.

Table 4
Comparison of our late fusion ResNet based approach (overall accuracy (OA) in %) with the best mid-level method: SIFT descriptors with Improved Fisher Vector (IFK-SIFT) encoding and the existing high-level deep methods: CafeeNet, VGG-VD-16 and GoogleNet on UC-Merced, WHU-RS19, RSSCN7 and AID datasets. Our approach provides consistently improved accuracy compared to both mid-level method and high-level deep methods on all datasets.

Method	UC-Merced (50%)	UC-Merced (80%)	WHU-RS19 (40%)	WHU-RS19 (60%)	RSSCN7 (20%)	RSSCN7 (50%)	AID (20%)	AID (50%)
IFK-SIFT	78.74 ± 1.65	83.02 ± 2.19	83.35 ± 1.19	87.42 ± 1.59	81.08 ± 1.21	85.09 ± 0.93	71.92 ± 0.41	78.99 ± 0.48
CaffeNet	93.98 ± 0.67	95.02 ± 0.81	95.11 ± 1.20	96.24 ± 0.56	85.57 ± 0.95	88.25 ± 0.62	86.86 ± 0.47	89.53 ± 0.31
VGG-VD-16	94.14 ± 0.69	95.21 ± 1.20	95.44 ± 0.60	96.05 ± 0.91	83.98 ± 0.87	87.18 ± 0.94	86.59 ± 0.29	89.64 ± 0.36
GoogleNet	92.70 ± 0.60	94.31 ± 0.89	93.12 ± 0.82	94.71 ± 1.33	82.55 ± 1.11	85.84 ± 0.92	83.44 ± 0.40	86.39 ± 0.55
Ours	**96.91 ± 0.36**	**97.72 ± 0.54**	**98.48 ± 0.37**	**98.88 ± 0.49**	**92.45 ± 0.45**	**94.0 ± 0.57**	**93.81 ± 0.12**	**95.73 ± 0.16**

Table 5
Comparison (overall accuracy in %) with the state-of-the-art approaches. Our approach provides a consistent improvement over the state-of-the-art on three datasets. Most notably a significant gain of 6.1% is obtained, compared to the state-of-the-art, on the large scale AID dataset. Note that on the WHU-RS19 dataset, Deep CNN Transfer (Scenario II) (Hu et al., 2015b) achieves 98.64% by employing VLAD encoding on the Conv layer features from the VGG-VD16. On the other hand, we do not employ any encoding scheme with the deep network.

Method	UC-Merced	WHU-RS19	RSSCN7	AID
BOVW + spatial co-occurrence kernel (Yang and Newsam, 2010)	77.70	–	–	–
Color Gabor (Yang and Newsam, 2010)	80.50	–	–	–
SPCK + SPM (Yang and Newsam, 2011)	77.40	–	–	–
Structural texture similarity (Risojevic and Babic, 2011)	86.0	–	–	–
Wavelet BOVW (Zhao et al., 2014a)	87.40	–	–	–
Unsupervised feature learning (Cheriyadat, 2014)	81.10	–	–	–
Saliency-guided feature learning (Zhang et al., 2015)	82.70	–	–	–
Concentric circle-structured BOVW (Zhao et al., 2014b)	86.60	–	–	–
Multifeature concatenation (Shao et al., 2013)	89.50	–	–	–
Pyramid-of-spatial-relations (Chen and Tian, 2015b)	89.10	–	–	–
CLBP (Chen et al., 2016)	85.50	–	–	–
MS-CLBP (Chen et al., 2016)	90.60	–	–	–
HHCV (Wu et al., 2016a)	91.80	–	86.40	–
DBN based feature selection (Zou et al., 2015)	–	–	77.0	–
Dirichlet (Kobayashi, 2014)	92.80	–	–	–
VLAT (Negrel et al., 2014)	94.30	–	–	–
Deep CNN Transfer (Scenario I: FC features) (Hu et al., 2015b)	96.88	96.71	–	–
Deep CNN Transfer (Scenario II: Conv features) (Hu et al., 2015b)	96.90	**98.64**	–	–
Deep filter banks (Wu et al., 2016b)	92.70	–	90.40	–
Class-specific codebook + two-step classification (Yan et al., 2017)	93.80	93.70	–	–
CaffeNet (Xia et al., 2017)	95.02	94.80	88.25	89.53
VGG-VD-16 (Xia et al., 2017)	95.21	95.10	87.18	89.64
GoogleNet (Xia et al., 2017)	94.31	92.92	85.84	86.39
This paper	**97.72**	98.20	**94.0**	**95.70**

In the seminal work of Xia et al. (2017), it was shown that among different mid-level methods, the SIFT descriptors with the Improved Fisher Vector (IFK-SIFT) encoding provide improved results for remote sensing scene classification. Table 4 shows the comparison of our late fusion two-stream deep ResNet architecture with the best mid-level method: IFK-SIFT and several existing high-level deep methods: the shallow CaffeNet and the very deep VGG-VD-16 and GoogleNet. All the baseline results are taken from Xia et al. (2017).The high-level deep feature approaches obtain consistently improved performance compared to the best mid-level method IFK-SIFT. Despite having only 8 layers, CaffeNet achieves competitive performance compared to very deep VGG-VD-16 and GoogleNet. Our late fusion based two-stream ResNet architecture provides consistent gain in performance compared to the existing high-level deep methods on all four datasets. In particular, a large gain in performance is achieved on the RSSCN7 and AID datasets. On the RSSCN7 dataset (20:80 training and test set ratio), the best mid-level method (IFK-SIFT) yields a mean recognition rate of 81.08%. The existing high-level deep methods: CaffeNet, VGG-VD-16 and GoogleNet provide mean classification scores of 85.57%, 83.98% and 82.55% respectively. Our approach achieves a mean classification rate of 92.45% outperforming best existing deep feature methods. A similar gain of 5.8% in mean accuracy is achieved, compared to best existing method, with 50:50 training and test set ratio on this dataset. On the recently introduced AID dataset (20:80 training and test set ratio), the best mid-level method (IFK-SIFT) provides a mean recognition rate of 71.92%. The existing high-level deep methods: CaffeNet, VGG-VD-16 and GoogleNet provide mean recognition rate of 86.86%, 86.59% and 83.44% respectively. Our approach provides superior performance compared to existing methods. Furthermore, a gain of 6.1% is obtained compared to the best existing deep feature method, with the 50:50 training and test set ratio on this dataset.

4.5. State-of-the-art comparison

Finally, we provide a comparison with the state-of-the-art approaches in literature. Our final image representation is late fusion two-stream ResNet architecture. Table 5 shows the comparison with the state-of-the-art methods in literature. We follow the same sampling setting as (Yang and Newsam, 2010; Sheng et al., 2012; Hu et al., 2015b) for fair comparisons, by taking 80 samples per class for training in case of the UC-Merced and 30 samples per class for training in case of the WHU-RS19 dataset. In case of the RSSCN7 and AID datasets, we use 50 training samples per class for training. On the UC-Merced dataset, the approach of Yang and Newsam (2010) integrating the spatial co-occurrence kernel within the bag-of-visual-words (BOVW) framework achieves a mean recognition rate of 77.7%. They also investigate integrating color information within Gabor features leading to a mean accuracy of 80.5%. The work of Yang and Newsam (2011) obtains a classification accuracy of 77.4% with a spatial pyramid co-occurrence based image representation that accounts for both photometric and geometric aspects of an image. Several approaches (Risojevic and Babic, 2011; Chen et al., 2016; Zhao et al., 2014a) aim to exploit texture information. Among these approaches, the multi-scale completed LBP feature provides superior performance with a mean recognition rate of 90.6%. A considerable gain in performance on this dataset can be observed with the use of deep feature based methods. The deep filter banks based approach of Wu et al. (2016b) achieves an accuracy of 92.7%. Transferring deep CNN features from the FC layer of the deep network (Deep CNN Transfer Scenario I: FC features) (Hu et al., 2015b) obtains a mean classification accuracy of 96.88%. Transferring deep CNNs from the Convolutional layers of the deep network encoded with the VLAD scheme (Scenario II: Conv features) achieves a recognition rate of 96.90%. Our approach achieves improved results (97.72%) on this dataset.

On the WHU-RS19 dataset, the recently introduced improved class-specific codebook using kernel collaborative representation based classification framework (Yan et al., 2017) achieves a mean accuracy of 93.7%. The CaffeNet and the very deep VGG-VD-16 and GoogleNet provide mean recognition rates of 94.8%, 95.1% and 92.9% respectively. Transferring deep CNN features from the FC layer of the deep network (Hu et al., 2015b) obtains a mean classification accuracy of 96.71%. Our approach achieves favorable results compared to existing methods. On this dataset, the best results (98.6%) are obtained when transferring deep CNNs from the Convolutional layers of the deep network encoded using the VLAD scheme. It is worthy to mention that our approach is complementary to (Scenario II: Conv features) method (Hu et al., 2015b) and combining the two approaches can be expected to provide further gain in the classification performance.

On the RSSCN7 dataset, the deep learning based feature selection approach (DBN) (Zou et al., 2015) achieves a mean recognition rate of 77.0%. The hierarchical coding vectors based classification approach (Wu et al., 2016a) achieves a classification result of 86.4%. The deep filter banks approach (Wu et al., 2016b) provides a classification performance of 90.4%. Our approach outperforms the best existing method (deep filter banks) with a mean classification accuracy of 94.0%. Finally, on the recently introduced AID dataset, the CaffeNet and the very deep VGG-VD-16 and GoogleNet methods provide mean recognition rates of 89.5%, 89.6% and 86.4% respectively. Our approach achieves the best results on this dataset with a mean classification accuracy of 95.7%.

5. Conclusions

In this paper, we address the problem of learning robust texture description within deep learning architectures for texture recognition and remote sensing scene classification. We design deep models by constructing a two-stream deep architecture where texture coded mapped images are used as a second stream and fuse it with the standard RGB stream. Furthermore, we investigate two fusion strategies, early and late fusion, to combine RGB and texture streams in our two-stream deep architecture. Experiments are conducted on several benchmark texture recognition and remote sensing scene classification datasets. Our results clearly demonstrate that the proposed late fusion two-stream deep architecture always improves the overall performance compared to the standard RGB stream deep network architecture for both recognition tasks. Further, our final combination leads to improved results compared to the state-of-the-art for remote sensing scene classification. In this paper, we investigate Local Binary Patterns (LBP) encoded CNNs and different deep network fusion architectures for texture recognition and remote sensing scene classification. Future work involves investigating alternative texture description techniques and fusion strategies for texture coded deep CNNs. Another future direction is to include training and testing the proposed approach on actual full-sized satellite images containing all available spectral bands besides RGB (e.g. Near Infrared).

Acknowledgements

This work has been funded by the Spanish project TIN2016-79717-R, the CHISTERA project M2CR (PCIN2015-251), SSF through a grant for the project SymbiCloud, VR starting grant (2016-05543), through the Strategic Area for ICT research ELLIIT, CENIIT grant (18.14), the project AIROBEST (317387, 317388) funded by the Academy of Finland, the project MegaMrt2, funded by the Electronic Component Systems for European Leadership (ECSEL) Joint Undertaking (grant agreement No. 737494) of the Horizon 2020 European Union funding programme. We acknowledge the computational resources provided by the Aalto Science-IT project and CSC IT Center for Science, Finland. We also acknowledge the computational support from Nvidia and the NSC.

References

Ahonen, T., Hadid, A., Pietikainen, M., 2004. Face recognition with local binary patterns. In: ECCV.

Anwer, R.M., Khan, F.S., van de Weijer, J., Laaksonen, J., 2017. Tex-nets: binary patterns encoded convolutional neural networks for texture recognition. In: ICMR.

Azizpour, H., Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPRW.

Borg, I., Groenen, F., 2005. Modern Multidimensional Scaling: Theory and Applications. Springer.

Bruna, J., Mallat, S., 2013. Invariant scattering convolution networks. TSE 35 (8), 1872–1886.

Caputo, B., Hayman, E., Mallikarjuna, P., 2005. Class-specific material categorisation. In: ICCV.

Chan, T.-H., Jia, K., Gao, S., Ma, Y., 2014. Pcanet: a simple deep learning baseline for image classification? TIP 24 (12), 5017–5032.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: delving deep into convolutional nets. In: BMVC.

Chen, S., Tian, Y., 2015a. Pyramid of spatial relations for scene-level land use classification. TGRS 53 (4), 1947–1957.

Chen, S., Tian, Y., 2015b. Pyramid of spatial relations for scene-level land use classification. TGRS 53 (4), 1947–1957.

Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W., 2010. Wld: a robust local image descriptor. PAMI 32 (9), 1705–1720.

Chen, L., Yang, W., Xu, K., Xu, T., 2011. Evaluation of local features for scene classification using VHR satellite images. In: JURSE.

Chen, X., Xiang, S., Liu, C.-L., Pan, C.-H., 2014a. Vehicle detection in satellite images by hybrid deep convolutional neural networks. LGRS 11 (10), 1797–1801.

Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y., 2014b. Deep learning-based classification of hyperspectral data. JSTARS 7 (6), 2094–2107.

Chen, X., Fang, T., Huo, H., Li, D., 2015a. Measuring the effectiveness of various features for thematic information extraction from very high resolution remote sensing imagery. TGRS 53 (9), 4837–4851.

Chen, Y., Zhao, X., Jia, X., 2015b. Spectral-spatial classification of hyperspectral data based on deep belief network. JSTARS 8 (6), 2381–2392.

Chen, C., Zhang, B., Su, H., Li, W., Wang, L., 2016. Land-use scene classification using multi-scale completed local binary patterns. SIVP 4, 745–752.

Cheriyadat, A., 2014. Unsupervised feature learning for aerial scene classification. TGRS 52 (1), 439–451.

Cheron, G., Laptev, I., Schmid, C., 2015. P-cnn: pose-based cnn features for action recognition. In: ICCV.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A., 2014. Describing textures in the wild. In: CVPR.

Cimpoi, M., Maji, S., Kokkinos, I., Vedaldi, A., 2016. Deep filter banks for texture recognition, description, and segmentation. IJCV 118 (1), 65–94.

Csurka, G., Bray, C., Dance, C., Fan, L., 2004. Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.-F., 2009. Imagenet: a large-scale hierarchical image database. In: CVPR.

dos Santos, J.A., Penatti, O.A.B., da Silva Torres, R., 2010. Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. In: VISAPP.

Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W., Multimodal deep learning for robust RGB-D object recognition. In: IROS.

Fathi, A., Nilchi, A., 2012. Noise tolerant local binary pattern operator for efficient texture analysis. PRL 33 (9), 1093–1100.

Feichtenhofer, C., Pinz, A., Zisserman, A., 2016. Convolutional two-stream network fusion for video action recognition. In: CVPR.

Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP.

Geng, J., Fan, J., Wang, H., Ma, X., Li, B., Chen, F., 2015. High-resolution SAR image classification via deep convolutional autoencoders. LGRS 12 (11), 2351–2355.

Guo, Z., Zhang, L., Zhang, D., 2010a. Rotation invariant texture classification using LBP variance (LBPV) with global matching. PR 43 (3), 706–719.

Guo, Z., Zhang, L., Zhang, D., 2010b. A completed modeling of local binary pattern operator for texture classification. TIP 19 (6), 1657–1663.

Guo, Z., Zhang, L., Zhang, D., 2010c. A completed modeling of local binary pattern operator for texture classification. TIP 19 (6), 1657–1663.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. In: CVPR.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR.

Hoffman, J., Gupta, S., Darrell, T., 2016. Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR.

Hu, F., Xia, G.-S., Wang, Z., Huang, X., Zhang, L., Sun, H., 2015a. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. JSTARS 8 (5), 2015–2030.

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015b. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Rem. Sens. 7 (11), 680–707.

Khan, F.S., van de Weijer, J., Anwer, R.M., Felsberg, M., Gatta, C., 2014. Semantic pyramids for gender and action recognition. TIP 23 (8), 3633–3645.

Khan, F.S., Anwer, R.M., van de Weijer, J., Felsberg, M., Laaksonen, J., 2015. Compact color-texture description for texture classification. PRL 51, 16–22.

Kobayashi, T., 2014. Dirichlet-based histogram feature transform for image classification. In: CVPR.

Kusumaningrum, R., Wei, H., Manurung, R., Murni, A., 2014. Integrated visual vocabulary in latent Dirichlet allocation-based scene classification for ikonos image. JARS 8 (1), 083690.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Handwritten digit recognition with a back-propagation network. In: NIPS.

Leung, T., Malik, J., 1996. Detecting, localizing and grouping repeated scene elements from an image. In: ECCV.

Leung, Thomas, Malik, J., 2001. Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV 43 (1), 29–44.

Levi, G., Hassner, T., 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: ICMI.

Liu, L., Zhao, L., Long, Y., Fieguth, P., 2012. Extended local binary patterns for texture classification. IMAVIS 30 (2), 86–99.

Liu, L., Fieguth, P., Wang, X., Pietikainen, M., Hu, D., 2016. Evaluation of LBP and deep texture descriptors with a new robustness benchmark. In: ECCV.

Liu, L., Fieguth, P., Guo, Y., Wang, X., Pietikainen, M., 2017. Local binary features for texture classification: taxonomy and experimental study. PR 62, 135–160.

Luo, B., Jiang, S., Zhang, L., 2013. Indexing of remote sensing images with different resolutions by multiple features. JSTARS 6 (4), 1899–1912.

Maenpaa, T., Pietikainen, M., 2004. Classification with color and texture: jointly or separately? PR 37 (8), 1629–1640.

Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep learning earth observation classification using imagenet pretrained networks. LGRS 13 (1), 105–109.

Matthieu Molinier, T.H., Laaksonen, Jorma, 2007. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. TGRS 45 (4), 861–874.

Negrel, R., Picard, D., Gosselin, P.-H., 2014. Evaluation of second-order visual features for land-use classification. In: CBMIW.

Ojala, T., Pietikainen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. PR 29 (1), 51–59.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24 (7), 971–987.

Ojansivu, V., Rahtu, E., Heikkila, J., 2009. Rotation invariant local phase quantization for blur insensitive texture analysis. In: ICPR.

Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR.

Penatti, O., Nogueira, K., Santos, J., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: CVPRW.

Perronnin, F., Dance, C., 2007. Fisher kernels on visual vocabularies for image categorization. In: CVPR.

Pietikainen, M., Zhao, G. Two Decades of Local Binary Patterns: A Survey. Available from: <arXiv:1612.06795>.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016. Generative adversarial text to image synthesis. In: ICML.

Risojevic, V., Babic, Z., 2011. Aerial image classification using structural texture similarity. In: ISSPIT.

Romero, A., Gatta, C., Camps-Valls, G., 2016. Unsupervised deep feature extraction for remote sensing image classification. TGRS 54 (3), 1349–1362.

Rubner, Y., Tomasi, C., Guibas, L., 2000. The earth mover's distance as a metric for image retrieval. IJCV 40 (2), 99–121.

Shao, W., Yang, W., Xia, G.-S., Liu, G., 2013. A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization. In: ICCVS.

Sharan, L., Liu, C., Rosenholtz, R., Adelson, E., 2013. Recognizing materials using perceptually inspired features. IJCV 103 (3), 348–371.

Sharma, G., ul Hussain, S., Jurie, F., 2012. Local higher-order statistics (LHS) for texture categorization and facial analysis. In: ECCV.

Sheng, G., Yang, W., Xu, T., Sun, H., 2012. High-resolution satellite scene classification using a sparse coding based multiple feature combination. IJRS 33 (8), 2395–2412.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In: NIPS.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: ICLR.

Tan, X., Triggs, B., 2007. Fusing gabor and LBP feature sets for kernel-based face recognition. In: AMFG.

Tan, X., Triggs, B., 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. TIP 19 (9), 1635–1650.

Tuia, D., Flamary, R., Courty, N., 2015. Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions. JPRS 105, 272–285.

ul Hussain, S., Triggs, B., 2012. Visual recognition using local quantized patterns. In: ECCV.

Varma, M., Zisserman, A., 2010. A statistical approach to texture classification from single images. IJCV 32 (9), 1705–1720.

Vedaldi, A., Lenc, K., 2015. Matconvnet: convolutional neural networks for matlab. In: ACM Multimedia.

Wang, X., Han, T., Yan, S., 2009. An HOG-LBP human detector with partial occlusion handling. In: ICCV.

Wu, H., Liu, B., Su, W., Sun, J., 2016a. Hierarchical coding vectors for scene level land-use classification. Rem. Sens. 8 (5), 436–453.

Wu, H., Liu, B., Su, W., Zhang, W., Sun, J., 2016b. Deep filter banks for land-use scene classification. LGRS 13 (12), 1895–1899.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., 2017. Aid: a benchmark dataset for performance evaluation of aerial scene classification. TGRS 55 (7), 3965–3981.

Yan, L., Zhu, R., Mo, N., Liu, Y., 2017. Improved class-specific codebook with two-step classification for scene-level classification of high resolution remote sensing images. Rem. Sens. 9 (3), 223–247.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: GIS.

Yang, Yi, Newsam, S., 2011. Spatial pyramid co-occurrence for image classification. In: ICCV.

Yang, Y., Newsam, S., 2013. Geographic image retrieval using local invariant features. TGRS 51 (2), 818–832.

Ylioinas, J., Hadid, A., Guo, Y., Pietikainen, M., 2012. Efficient image appearance description using dense sampling based local binary patterns. In: ACCV.

Ylioinas, J., Hong, X., Pietikainen, M., 2013. Constructing local binary pattern statistics by soft voting. In: SCIA.

Zeiler, M., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: ECCV.

Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C., 2007. Local features and kernels for classification of texture and object categories: a comprehensive study. IJCV 73 (2), 213–218.

Zhang, J., Huang, K., Yu, Y., Tan, T., 2011. Boosted local structured HOG-LBP for object localization. In: CVPR.

Zhang, F., Du, B., Zhang, L., 2015. Saliency-guided unsupervised feature learning for scene classification. TGRS 53 (4), 2175–2184.

Zhao, L., Tang, P., Huo, L., 2014a. A 2-d wavelet decomposition-based bag-of-visual-words model for land-use scene classification. IJRS 35, 2296–2310.

Zhao, L., Tang, P., Huo, L., 2014b. Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. JSTARS 7 (12), 4620–4631.

Zhong, Y., Zhu, Q., Zhang, L., 2015. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. TGRS 53 (11), 6207–6222.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. LGRS 12 (11), 2321–2325.