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

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/

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 fine-tuning or ensemble of RGB network architectures, for remote sensing scene classification.