Sky-CNN: A CNN-based Learning Approach for Skyline Scene Understanding

Ameni Sassi, Wael Ouarda, Chokri Ben Amar, Serge Miguet

▶ **To cite this version:**

Ameni Sassi, Wael Ouarda, Chokri Ben Amar, Serge Miguet. Sky-CNN: A CNN-based Learning Approach for Skyline Scene Understanding. International Journal of Intelligent Systems and Applications, 2019, 4, pp.14 - 25. 10.5815/ijisa.2019.04.02 . halshs-02471883

HAL Id: halshs-02471883
https://shs.hal.science/halshs-02471883v1

Submitted on 9 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

I.J. Intelligent Systems and Applications, **2019, 4, 14-25**
Published Online April 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2019.04.02

Sky-CNN: A CNN-based Learning Approach for Skyline Scene Understanding

Ameni Sassi
REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax, ENIS, BP 1173, 3038, Sfax, Tunisia
LIRIS, Université de Lyon, UMR CNRS 5202, Université Lumière Lyon 2, 5 av. Mendès-France, Bât C, N 123, 69676. Bron, Lyon, France
E-mail: ameni.sessi.tn@ieee.org

Wael Ouarda, Chokri Ben Amar
REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax, ENIS, BP 1173, 3038, Sfax, Tunisia
E-mail: {wael.ouarda, chokri.benamar}@ieee.org

Serge Miguet
LIRIS, Université de Lyon, UMR CNRS 5202, Université Lumière Lyon 2, 5 av. Mendès-France, Bât C, N 123, 69676. Bron, Lyon, France
E-mail: serge.miguet@univ-lyon2.fr

Received: 16 October 2018; Accepted: 13 December 2018; Published: 08 April 2019

Abstract**—**Skyline scenes are a scientific matter of interest for some geographers and urbanists. These scenes have not been well-handled in computer vision tasks. Understanding the context of a skyline scene could refer to approaches based on hand-crafted features combined with linear classifiers; which are somewhat side-lined in favor of the Convolutional Neural Networks based approaches. In this paper, we proposed a new CNN learning approach to categorize skyline scenes. The proposed model requires a pre-processing step enhancing the deep-learned features and the training time. To evaluate our suggested system; we constructed the SKYLINEScene database. This new DB contains 2000 images of urban and rural landscape scenes with a skyline view. In order to examine the performance of our Sky-CNN system, many fair comparisons were carried out using well-known CNN architectures and the SKYLINEScene DB for tests. Our approach shows it robustness in Skyline context understanding and outperforms the hand-crafted approaches based on global and local features.

Index Terms**—**Convolutional Neural Network, deep learning, scene categorization, skyline, features representation, deep learned features.

I. Introduction

Scene Understanding is a well-discussed issue in many application domains. Low-level handcrafted features have been widely used for scene description and classification where scene images could be characterized globally or by local signatures [1]. Recent trends pursue to learn scene features automatically using powerful deep learning models such as convolutional neural networks (CNN). These networks, in particular, provide a powerful end-to-end framework which tightly integrates feature extraction and classification to performance in many challenging computer vision tasks. The features learned by deep Convolutional Neural Networks ranging from low-level to high-level representations in the hidden layers have also prompt different researchers to investigate the ways to take advantage of these architectures. Since Convolutional Neural Networks brighten in diverse computer vision applications, deeper architectures have been proposed to reach better results. These networks may require a huge number of parameters often in the range of millions and could expose the "vanishing gradients" and the over-fitting problems because of its higher depth. Thus, the choice of the most appropriate architecture is challenging.

Skyline scenes are very particular scenes that have not been well-handled on image understanding. These scenes were scientifically interesting for geography and urbanism domains [2, 3, 4, 5, 6] since they are considered as a particular dimension of some cities. Our skyline scenes understanding mission was lead in a multidisciplinary research project named SKYLINE [1]. Citizens were implicated in this project, as they could evaluate their skylines and interact to predict the future of their cities views. Thus, our objective was to understand the category of a landscape scene depending on its component elements, in order to classify each scene into urban or rural skyline. To ensure a legitimate categorization of this type of scenes, we need a database containing an interesting diverse collection of skyline

[1] The SKYLINE project website: http://recherche.univ-lyon2.fr/skyline/wordpress/?page_id=98

Copyright © 2019 MECS

photographs. To the best of our knowledge, referencing to our scientific readings, there is no specific database for skylines with a significant number of images. The only database we have found was the Skyline-12 dataset [7] with 120 high-resolution skyline images from twelve cities. The creation of such database, with an interesting number of diverse skyline scenes, will enable researchers working on skyline-based image processing tasks to develop and evaluate their algorithms. A Skyline scenes database would be a collection of very specific scenes with respect to the geometric definition of skylines, so it would be interesting to propose an adequate deep architecture for characterizing these specific scenes.

Within this work, we train a new CNN architecture to categorize skyline scenes into two contexts: urban and rural; taking into consideration the relevant parts in these scene images. All skyline scenes share a very particular part that is the sky. The horizon line separating the sky from ground objects puts on show the distinctive part under-skyline containing the relevant information about the scene. To get a deeper representation of these scenes, we adopted the residual connections in our proposed convolutional neural network architecture to combat the problems raised with deep neural networks.

The remainder of the paper is structured as follows: the first section exposes some related works from the state-of-the-art. Then, in the second part, we will describe our proposed CNN model for skyline scenes categorization. Afterwards, we present all the details about the new SKYLINEScene database. The next section presents the experimental settings, the results and the discussion handled within our proposed approach evaluation. The last section covers the conclusion of the realized works.

II. Related Works

Computer vision has made a prodigious progress on image understanding, notably on classification tasks. The hand-engineered features with linear classifiers have been the keywords to describe and classify scene images, for a long time. A variety of visual features and classifiers are used and adapted depending on the application. Static scenes could be represented by numerous local and global features. Recently, Convolutional Neural Networks have greatly advanced the performance in these tasks. Deep features derived from CNNs have fared quite good on object extraction [8] and recognition [9, 10] as well as on scene classification [11, 11, 13].

Between global and local descriptors, each work selects the appropriate visual features to represent images for classification purposes. The Scale Invariant Feature Transform (SIFT) algorithm and all of its derivatives [14, 15] were applied as local descriptors for scene categorization. The Histogram of Oriented Gradients (HOG) is an image-cell-level local descriptor that proposed in the literature. The work in [16] proposes a comparison of this latter descriptor with SIFT and GIST [17] descriptors. The HOG-based description leads to a higher classification rates than the ones obtained with other descriptor using the SUN397 dataset. The local descriptor developed mainly for texture classification algorithms [18], was entitled the Local Binary Pattern (LBP). The LBP was even used for the large-scale scene categorization task in the previously mentioned related work [18] and adopted for a real-time classification of landscape scenes [19] by reason of its low-computational cost.

Global descriptors summarize statics about the integral image without proceeding the segmentation into regions or the detection of some objects. There are several methods to globally describe a scene image such as the histograms, the Principal Component Analysis (PCA) [20], the Fisher Vector (FV) [14] and the Bag of Words (BoW) models. Some global features were built by means of a combination of diverse methods. In this work [21], the histograms of wavelet texture and quantized color were used as global descriptors for indoor/outdoor scene classification. A combination of two other types of histograms, which are the edge direction histogram and the color histograms, were selected for City vs. Landscape images classification [22]. In the bag of words algorithms, instead of directly applying the global visual features, an intermediate representation of the image is proposed. First, the visual words dictionary for scene images is created. Then, based on this dictionary, the bag-of-words models are built to represent scenes. A sample work using the BoW algorithm [23] is proposed for a hybrid holistic/semantic approach for scene classification. Using the Hierarchical Matching Pursuit (HMP) to learn holistic features and the Semantic Spatial Pyramid (SSP) to represent the spatial object information, this previous work combined these two strategies for images representation with support vector machine SVM classifier to propose a scene classification methodology. Their hybrid approach reached a global accuracy of 78.2% using a dataset of 700 images containing six natural scenes (forests, coasts, rivers/lakes, plains, mountains, and sky/clouds). Another work [19] dedicated to landscapes scenes recognition compared the use of Bag of Words methods with others in terms of classification accuracies and execution time.

Over the past few years, the convolutional neural networks (CNN) have gained an important research interest for the field of image understanding and categorization. These deep learning architectures were first developed by [24]. The work in [13] proposed a CNN architecture with some specific parameters for image classification. Their developed architecture achieved a rate of 62.5% for the top-1 class on the ImageNet benchmark containing 1000 object categories for 1.2 million images. The proposed AlexNet model in [13] was compared in this work [11] with other convolutional architectures. The classification accuracy reached was 59.85% for the MIT indoor dataset and 43.74% for the SUN 397 dataset. The C-RNN architecture proposed in [11] is trained in an end-to-end manner from raw pixel images. CNN layers are first processed to generate middle-level features. RNN layer is then learned to encode spatial dependencies. The experiments achieved in this work proved that the C-

Copyright © 2019 MECS

RNN can learn better image representation, especially for images with obvious spatial contextual dependencies. Fine-tuning the C-RNN, [11] achieved 51.14% classification accuracy on SUN397, and 65.07% on the MIT-indoor dataset. Using the same classification datasets, the MOP-CNN model proposed in [12] achieved 51.98% and 68.88% on the SUN397 and MIT-indoor respectively. The scheme called Multi-scale Orderless Pooling (MOP-CNN) extracts CNN activations for local patches at multiple scale levels, performs orderless vectors of locally aggregated descriptor pooling of these activations at each level separately, and concatenates the result.

Discussion

To succeed scene classification tasks, we have to ensure a good representation of the scene image. The papers cited in the previous related work, building the representation of scene images on local and global descriptors, try to use different types of characteristics: shape, texture, and color. An urban or rural Skyline scene image is usually characterized by its distinctive geometric aspects of the horizon line and the under-skyline region which can be easily described by a combination of color (green vegetation, gray buildings, etc.), texture and geometric information.

Most of the Deep Neural Networks models and the associated benchmarks are not specifically designed for urban and rural contexts. Classifying a Skyline scene to one of these generic contexts require a relevant choice of the network architecture and the target dataset. For these reasons, we have designed our specific deep architecture.

Regarding the related work of global and local hand-crafted features, we found a lack in using the geometric information to describe the scene images in [23]. The authors in this latter work have mainly used a texture-based representation combined with color information. The work of Xiao *et al*. [16] neglects both the color and the geometric features on describing scene images for classification, where the authors have focused on using texture descriptors. The works elaborated in [21] and [22] have combined the color and the texture features to represent scene images, while the geometric information have not been considered.

The Skyline scenes representation proposed in our latest work [8], was based mainly on geometric and texture hand-engineered features. In order to fuse the different types of information, we adopted a CNN features representation based essentially on the non-linear transformation of the under-skyline image using a cascading of multiple convolution layers; Pooling for dimensionality reduction and Rectified Linear Unit (ReLU) to include all non-linear patterns in the image. Three levels of features are represented throughout our architecture:

- low-level features like lines, corner and color presented at the beginning of the Sky-CNN (Fig. 9);
- middle-level features such as shape and edge extracted from the middle layers of the proposed architecture;
- and high-level features (texture) obtained by the combination of previous learned features by the classifier.

III. Proposed System

Our proposed system is based on deep learning methods to classify landscape images. To deal with this classification process, we have designed a deep neural network with residual connections within convolutional layers. The proposed architecture has been customized to perform well on skyline scenes classification. Given the input image, we processed it to obtain the resized under-skyline images for the three color channels (red, green, blue). In the next step, the residual networks are trained to ensure a deep learning with additional functions to be learned. After the training process, The features maps are flattened to set the final Fully Connected (FC) layer which is followed by a softmax layer to output the probabilities scores for our two desired classes: urban and rural. Fig. 1 illustrates the previous-detailed process. The proposed deep network ensures a customized deep representation of skyline scenes.

A. Skyline Image Preprocessing

The horizon line is a key feature in some landscape scenes. This border is often called the skyline. In the literature, many definitions of the skyline are proposed. In the Collins English Dictionary, the skyline is defined as the line at which the earth and sky appear to meet; or the outline of buildings, mountains, trees, or other elements seen against the sky. It is also considered as the line or shape that is formed where the sky meets buildings or the land. Another definition proposed by MacMILLAN Dictionary denotes that the skyline is the shapes made by buildings or mountains when you see them against the sky. Not so far from the last definition, according to the Oxford English Dictionary, the skyline is the line or shape that is formed where the sky meets buildings or the land. We can conclude from all these definitions that the skyline is the line separating the sky from the other elements in a landscape view. These elements can be buildings, Towers, mountains, trees or others, and they define the global shape of a skyline by their layout and their appearance from a specific point of view.

A Skyline scene is an image showing the global view characterizing a city or a rural place with the sky in the background. This image describes the layout of a landscape in front of the sky. For geographers, skylines expose some set of elements of the city structure where some buildings are more visible than others. Moreover, Skylines represent a view that comprised also trees, bridges, technical infrastructure, and topography.

The relevant region characterizing a landscape scene and helping to reveal the nature of that scene is the under-skyline area where we can apply different features extraction algorithms. The pre-step of analyzing this area

Copyright © 2019 MECS

should be the extraction of the horizon line separating the sky from ground elements. The algorithm adopted for the skyline extraction, presented in [6], contains four main steps. The first one is the application of the Canny filter to extract edges, only the most contrasted contours will be kept by the filter. From the obtained edge map, an upper envelope is extracted, which is a disconnected approximation of the skyline. A graph is then constructed and a shortest-path algorithm is used to link discontinuities. The obtained curve defining the skyline will help to get the under-Skyline image. The steps of the skyline extraction are represented in Fig. 2.

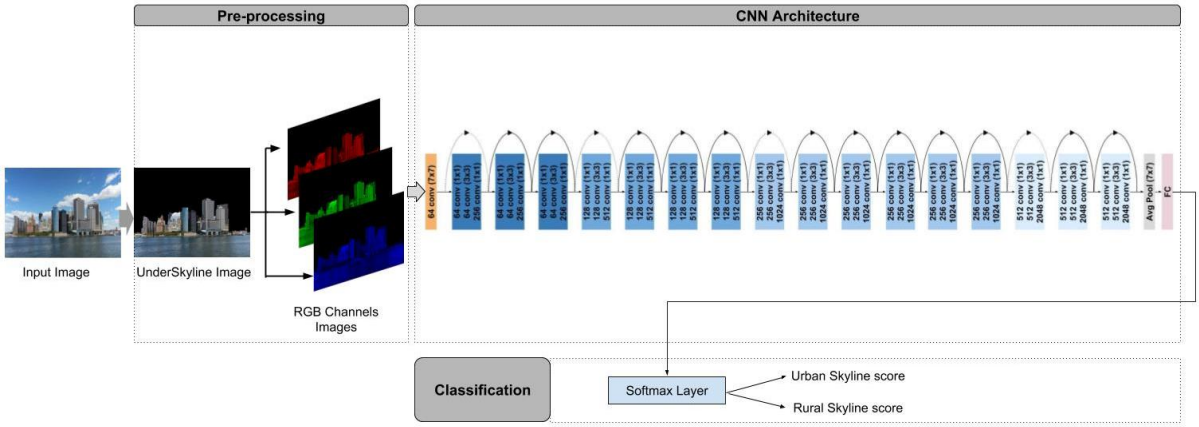

Fig.1. Architecture of the Sky-CNN System

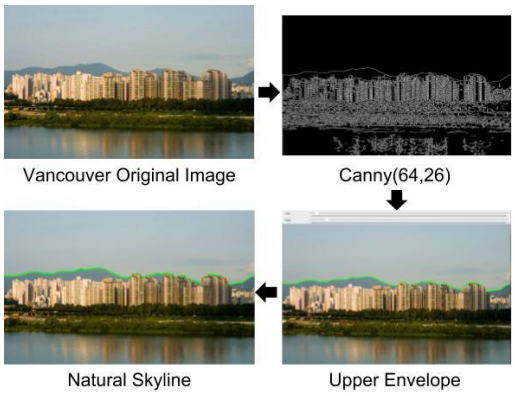

Fig.2. The skyline extraction steps [6]

The Canny edge detector is a multi-stage edge detector algorithm beginning with a noise reduction using a 5x5 Gaussian filter. Using Sobel kernel in both horizontal and vertical direction, we get the first derivative images G_x and G_y which are used to find the intensity gradient of the Skyline image represented by the edge gradient (1) and direction (2).

$$EdgeGradient = \sqrt{G_x^2 + G_y^2} \tag{1}$$

$$Angle(\theta) = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{2}$$

After computing the gradient magnitude and direction, the step of non-maximum suppression is established to remove the unwanted pixels which may not be relevant components of the edge. The last step to detect Canny edges is the thresholding where the decision about the real edges is made. Two threshold values are needed: minVal and maxVal. Using a Hysteresis thresholding with these values and considering "sure-edge" pixels, the image, pixels are classified into edges and non-edges. These minimum and maximum values help to specify which skyline to extract [6].

B. Deep Learning Architecture

The convolutional neural network architecture we have proposed is designed using the deep residual connections, which are first introduced in [25]. Deep ResNet is a type of specialized neural network that helps to handle more sophisticated deep learning tasks and models. It has received quite a bit of attention at recent IT conventions [26, 27, 28] and is being considered for helping with the training of deep networks. In deep learning networks, a residual learning framework helps to preserve good results through a network with many layers. One problem commonly cited by professionals is that with deep networks composed of many dozens of layers, accuracy can become saturated, and some degradation can occur. Some talk about a different problem called "vanishing gradient" [29, 30] which the gradient fluctuations become too small to be immediately useful. The deep residual network deals with some of these problems by using residual blocks, which take advantage of residual mapping to preserve inputs [32]. By utilizing deep residual learning frameworks, engineers can experiment with deeper networks that have specific training challenges.

The Deep Residual Networks with its different realizations obtained very successful results in the ImageNet and MS-COCO competition [25]. The chosen model for our CNN architecture was the ResNet50. The residual blocks focus on the problem of training a deep architecture by introducing identity skip connections in order to allow layers to copy their inputs to the next layer. As illustrated in Fig. 3, the general form of each residual unit can be expressed as:

Copyright © 2019 MECS

$$y_l = h(x_l) + \mathcal{F}(x_l, \mathcal{W}_l); x_{l+1} = f(y_l) \quad (3)$$

Where x_l and x_{l+1} represent the input and the output of the l^{th} unit, and $\mathcal{F}$ is a residual function. The identity mapping is expressed as $h(x_l) = x_l$ and f is a Rectified Linear Unit (ReLU) function [25]. $\mathcal{W}_l$ is a set of weights that are associated with the l^{th} Residual Unit. ResNet models fit a residual mapping to predict the parameters needed to reach the final prediction from one layer to the next, instead of fitting the latent weights to predict the final class at each layer. The identity mapping enables the model to bypass a typical CNN weight layer if the current layer is not necessary and to avoid overfitting to the training set. The intuitive idea behind the residual network architectures that it ensures that the next layer learns something new and different from what the input has already encoded. In addition, these residual connections help to raise the vanishing gradients problem.

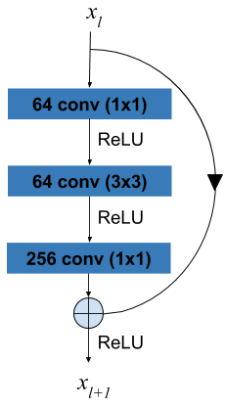

Fig.3. Residual unit diagram

Fig.4. Samples from SKYLINEScene Database: (a) Urban Skylines (b) Rural Skylines

C. Skyline Scenes Classification

The softmax classifier denoted as SMC is one of the commonly used classifiers. This classifier is the generalization of the binary logistic regression to multiple classes. We used such a classifier to differentiate between two categories of skylines: rural and urban skylines. The softmax regression allows us to manage k classes $t(i) \in \{1,..,k\}$ with a training set of m labeled examples $\{(z(1), t(1)), ..., (z(m), t(m))\}$ where the input features are $z^i \in \Re^n$. In our case, we were in a binary classification where the logistic regression took the form:

$$f_{w(3)}(z) = \frac{1}{1 + \exp\left(-W^{(3)T} z\right)} \quad (4)$$

where f_w is a sigmoid function having as parameters $W^{(3)}$, knowing that this non-linear function has already shown its robustness in previous work [8]. The input z of the SMC is the new representation of the skyline scene features learned by Sky-CNN system. The softmax layer computes the scores for each of urban and rural class. The Softmax's parameter $W^{(3)}$ is trained to minimize the cost function.

IV. SKYLINEScene Database

SKYLINEScene [2] database was essentially derived from photo storage and sharing websites, some dataset [7] and the photographs presented in the photo-questionnaire used in the SKYLINE project. The motivation behind the introduction of this new database is the deficiency of landscapes scenes datasets. The architecture, the infrastructure and the topography in diverse places around the world have plenty changed. The existing datasets are quite old and contain low-resolution images [21, 22]. The only database we have found was the Skyline-12 dataset [7] with 120 high-resolution skyline

[2] The SKYLINEScene database link: *https://goo.gl/6Yncre*

images from twelve cities. Our database is a wide extension of the Skyline-12.

The SKYLINEScene DB consists of 2000 true color images of skyline scenes, size-normalized to 320x240 pixels each. Our database contains 1000 images of urban skylines from various cities such as Shanghai, New York, Dubai, Paris, Hong Kong, Istanbul, etc., taken from different points of view in day or night time. The second part of the database contains natural landscapes images showing grasslands, valleys, deserts, forests, mountains and also some cities with natural skyline view. The creation of this database will enable researchers working on skyline-based image processing tasks to develop and evaluate their algorithms. Some samples images of the SKYLINEScene DB are shown in Fig. 4.

Besides the diversity assured in the cityscape photos, we have tried to gather a variety of natural landscapes photos in different seasons where the trees and the grass change their color, and the valleys or the mountains could be covered by the snow. The SKYLINEScene DB contains landscapes photos taken in autumn, marked with the orange-yellow colors of vegetation. There are also an interesting number of landscapes photographed in spring season. Snowy and rocky landscapes are part of our collected natural skyline scenes. The urban landscapes photos were picked up from almost 30 cities and divided almost equally into two parts: landscapes with water plan

or without. Since photographs tend to capture landscapes photos at different times and weather, our SKYLINEScene DB contains 204 urban skylines taken in sunset or night-time where the color of buildings and water plans changes depending on the sun position or the artificial lights conditions.

All the images collected respect the geographical definition of the skyline; they have to show the global view characterizing a city or a rural place with the sky in the background. In this way, we could utilize the geometric features of the horizon line that could be easily extracted. The SKYLINEScene database contains some landscapes photos with ambiguous skyline; its global view shows a mixture of artificial elements and natural components such as mountains or land field. Fig. 5 depicts some samples of these landscapes.

Fig.5. Samples of mixed skylines from SKYLINEScene DB

V. Experiments and Results

In this section, we evaluate the performance of our proposed Sky-CNN system compared to different deep architectures using the SKYLINEScene database for testing. The deep features representation resulting from our system is compared to some hand-crafted features representation. Several qualitative and quantitative evaluations have been also handled to highlight the contribution of the sky line information on describing landscapes scene.

A. Experimental Settings

The experiments were carried out using a Core i7 2-GHz machine. Our system implementation is written in a combination of C/C++ and Matlab languages. The training options of the Sky-CNN system are summarized in Table 1 and we used only a CPU based implementation. The SKYLINEScene database was divided randomly into 70% of samples for training and 30% for testing.

Table 1. Training Options for the Sky-CNN system

Parameter	Value
Mini-batch Size	10
Maximum number of epochs	1
Initial learning rate	0.0001
Frequency of network validation	3
Patience of validation stopping	Inf

B. Experimental Results

Recently, various CNN architectures have been proposed such as GoogLeNet [9], ResNet [25], Inception [31] and others. Our first experiments were applied to investigate the performance of some CNNs on classifying the collected images of SKYLINEScene database. These experiments will ensure a fair comparison between different deep representations on the skyline scene images providing the classification accuracies of urban and rural scenes. Each one of these neural networks has been trained on over a million images (ImageNet) and can classify images into 1000 object categories. The network has learned rich feature representations for a wide range of images. The pre-trained network will help to transfer the learned features in order to classify landscapes scenes.

The evaluation results for urban/rural landscapes classification are shown in Table 2. The obtained rates for the different CNN architectures were highly interesting especially for the pre-trained ResNets. Our Sky-CNN model outperforms other evaluated convolutional models with 99.33% classification accuracy for urban skyline scenes and a fully-accurate classification for rural landscapes scenes which give a rate of 99.67% as global accuracy. Besides this quantitative evaluation, we have examined the qualitative performance of the tested deep architectures by evaluating the specificity and the sensitivity of the classifier. The resulting ROC curves are shown in Fig. 6 where we can note that our proposed pre-trained model Sky-CNN not only challenges other architectures but also outperforms them on skyline scenes classification.

The next evaluations are about the performance of

Copyright © 2019 MECS

deep learning models compared to hand-engineered features on describing our skyline scenes for categorization. The global classification accuracy based only on the transformed geometric features [2] extracted from the horizon line reaches 85.8% [8] as illustrated in Table 3. For global image description, the classification rates were computed using the feature vector resulting from the RGB histograms and in this case the global accuracy was limited to 89.05%. Using a texture representation of landscapes scenes, the global classification accuracy reached is 62.14%. Combining these low-level features, we obtain a classification rate of 84.92%.

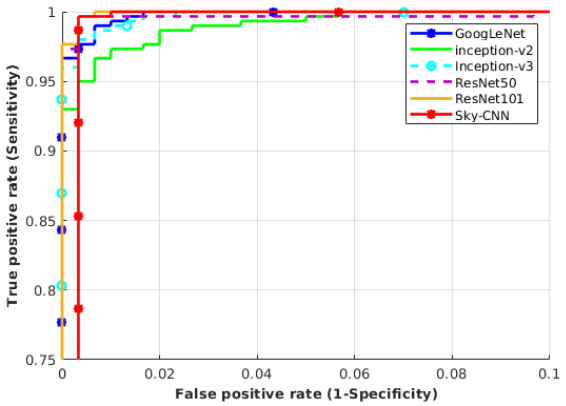

Fig.6. ROC curves for different CNN architectures

Table 2. Classification Accuracies (Acc.) depending on the deep training architecture

CNN Architecture	Global Acc.	Urban Acc.	Rural Acc.
GoogLeNet [9]	99.00%	98.67%	99.33%
InceptionResNet-v2 [31]	98.00%	98.67%	97.33%
Inception-v3 [31]	99.00%	98.33%	99.67%
ResNet-50 [33]	99.00%	99.33%	98.67%
ResNet-101 [25]	99.17%	98.33%	100.00%
Proposed Model: **Sky-CNN**	**99.67%**	**99.33%**	**100.00%**

Table 3. Classification accuracies depending on the description method

Features Vector	Global Acc.	Urban Acc.	Rural Acc.
Geometric [8] (Skyline)	85.80%	82.82%	88.73%
Color [22] (RGB Histograms)	89.05%	89.27%	88.82%
Texture [34] (Texton)	62.14%	68.20%	55.38%
Geometric+Color+Texture[8]	84.92%	84.92%	84.92%
Sky-CNN Deep learned features	**99.67%**	**99.33%**	**100.00%**

The confusion matrices in Fig. 7 illustrate the classification rates of urban and rural skyline scenes and also the number of true positives TP, true negatives TN, false positives FP, and false negatives FN. The matrices for urban and rural skylines classification obtained using some deep architectures are shown in Fig. 7 besides the one obtained for our proposed Sky-CNN (Fig. 7(a)). These matrices represent the confusion rates in percentages with the images number of positive and negative classes. The confusion matrix we got using the CNN architecture proposed in [9], which is based on transfer learning the GoogLeNet, is displayed in Fig. 7(f). Next to this last matrix, we present the ones displaying the classification rates based on fine-tuned different CNNs architectures proposed in the literature. These matrices are shown in Fig. 7.

C. Skylines pre-processing Evaluation

In order to emphasize the usefulness of the Skylines pre-processing step on representing landscape scenes, we compare, in this section, our Sky-CNN system to different CNN architectures using the SKYLINEScene database with and without proceeding to the pre-processing phase. The pre-processing of the Skyline scenes consists of considering only the part under the extracted horizon line, so the pixels representing the sky are neglected and appear as black pixels on the pre-processed images. The experiments are carried out to compare the classification rates and also the execution time obtained by testing different deep architectures for original and pre-processed Skyline scene images.

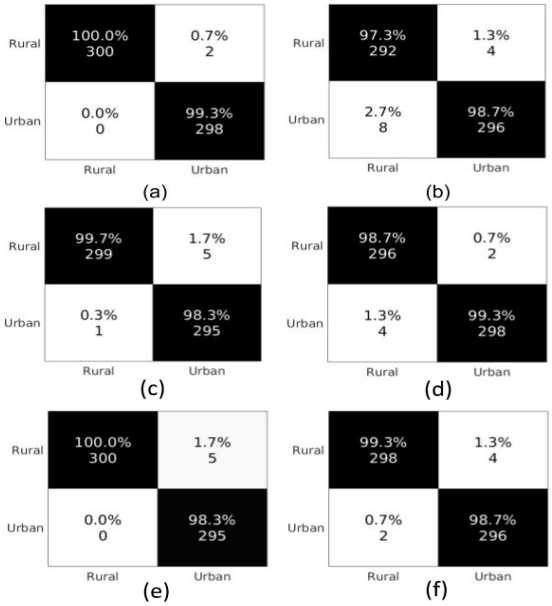

Fig.7. Confusion Matrices : (a) Sky-CNN (b) Inception-ResNet-v2 [31] (c) Inception-v3 [31] (d) ResNet-50 [33] (e) ResNet-101 [25] (f) GoogLeNet [9]

Table 4 shows the elapsed time (seconds) on the training process and also in the classification for both the whole Skyline images and the pre-processed Skylines using diverse architectures. The obtained numbers highlight the improvement; mostly for the training time, going through the pre-processing phase of landscapes images. Expect the GoogLeNet, the Sky-CNN spend the least training and testing time of all other CNNs architecture. The global classification rates illustrated by Table 5 enhance the improvement brought by the Skylines pre-processing step. The obtained accuracies using the under-Skyline images are better than the ones

Copyright © 2019 MECS

obtained using the whole Skyline images, for almost all the CNN architectures. Our proposed architecture outperforms all the other models on classifying the under-skyline images. The proposed Sky-CNN architecture has a complexity of $O(n^4)$ during the training phase; once the model is trained, the complexity of the test is logarithmic.

Table 4. Execution Time in seconds for some tested CNN architectures depending on the processed Skyline images

CNN Architecture	Whole Skyline Images		Pre-Processed Skyline Images	
	Train Time	Test Time	Train Time	Test Time
GoogLeNet [9]	19719	71.22	19619	69.35
Inception-ResNet-v2 [31]	147993	500.26	146663	461.37
Inception-v3 [31]	63767	209.10	63469	207.12
ResNet-101 [25]	79679	269.03	77396	262.28
Sky-CNN	**50799**	**165.38**	**46296**	**165.57**

Table 5. Global Accuracy rates for some tested CNN architectures depending on the processed Skyline images

CNN Architecture	Global Accuracy Rates	
	Whole Skyline Images	Pre-Processed Skyline Images
GoogLeNet [9]	99.00%	99.33%
Inception-ResNet-v2 [31]	98.00%	99.33%
Inception-v3 [31]	99.00%	99.00%
ResNet-101 [25]	99.17%	99.50%
Sky-CNN	**99.00%**	**99.67%**

Discussion

Analyzing these last experiments and the obtained results, we can affirm the authentic and the discriminant contribution of the proposed sky-CNN system on learning landscape scenes features. In fact, the specific pre-processing phase in the system adjusts the deep features learning process to provide a better representation of skyline scenes. The comparisons handled within Table 4 and Table 5 highlight the improvement provided by the pre-processing of the Skyline scenes in both time and accuracy. Therefore, the Sky-CNN combines the color, the geometric and the texture information related to the skyline area throughout the hidden layers of the model, discarding the useless information that the sky may contain. Our proposed deep system outperforms the hand-crafted previously-proposed approaches even when they combine diverse features. The Sky-CNN succeed on joining relevant features especially for Skyline scenes description. The activations features visualization of our proposed convolutional neural network are detailed in Section E.

D. Failure Modes

We further qualitatively analyze the failure modes of our proposed model and other deep classification models we have tested; to capture the effectiveness of the Sky-CNN model on classifying skyline scenes. The confusion matrices in Fig. 7 illustrate the number of the miss classified Skyline scenes for each tested architecture. The matrix associated to the proposed Sky-CNN system in Fig. 7(a); depicts that there are only two miss-classified urban skyline. Fig. 8(a) displays these two wrongly classified images. For urban landscapes, the miss-classification can be related to the horizon line area. Flat skylines, which is the case for all the urban miss-classified scenes for all the tested architecture. The cityscapes photos without apparent tall buildings are also confusing images. The skylines of some cities appear in a flat shape from some specific point of view or when the photo is taken from high position faraway, which is the case of the second and the fifth images in Fig. 8(c) for Paris and Madrid skylines. The mixed, natural and artificial, elements appearing in the horizon line are also ambiguous for the classification task that appear in Fig. 8(b)(c)(d)(f). The miss classified rural landscapes are mostly the ones showing mixed skylines or ambiguous colors. The rural landscapes with colors similar to artificial ones or taken in sunset time are confused with cityscapes. The first and the fifth rural landscapes in Fig. 8(b) depicting mountainous skylines with colors similar to artificial elements are confused with buildings. The appearance of buildings or artificial elements on rural skylines distorts the representation of that skyline and leads to the miss classification.

E. Sky-CNN Features Visualization

In this experimentation part of our work, we examine the activations of different layers of the proposed convolutional neural network. These experimentations help to discover the learned features by comparing areas of activation with the original input image. Through qualitative visualization and empirical analysis of the activations, we explored the purpose of the pre-processing step of skyline images and the relevance of the horizon line area on classifying skyline scenes.

To reveal the specificity of the features learned by our Sky-CNN model, we analyzed the activations of some layers for the miss-classified images. Our Sky-CNN model was compared to a fine-tuned ResNet-50. The failure modes obtained by fine-tuning the pre-trained ResNet-50 using the SKYLINEScene as a test database are shown in Fig. 8(d). Within this figure, an image of the Washington skyline showing the White House was miss-classified to a rural landscape. Analyzing the activations for this first scene, we began with Fig. 9 that depicts the strongest activations within the first convolutional layer of the pre-trained ResNet-50 for the whole Skyline image (Fig. 9(a)) and for the under-Skyline image (Fig. 9(a)). In this last figure, the horizon line and also the horizon edges below are distinctive with whiter pixels which implies that this under-skyline area is strongly positive activated. Going deeper, and exactly to the 26th convolutional layer, we investigated activations in some channels to reveal the pertinent learned parts of skyline scenes. Fig. 10 shows that the distinctive buildings appearing on the left and the right of the horizon line still highly activated for the Sky-CNN system. To investigate

Copyright © 2019 MECS

only positive activations, we visualize in the same figure the third channel activations of the 26th relu layer. These last figures prove that the learned features by our Sky-CNN model highlight the area around the buildings neglecting the information that the sky could contain such as the clouds.

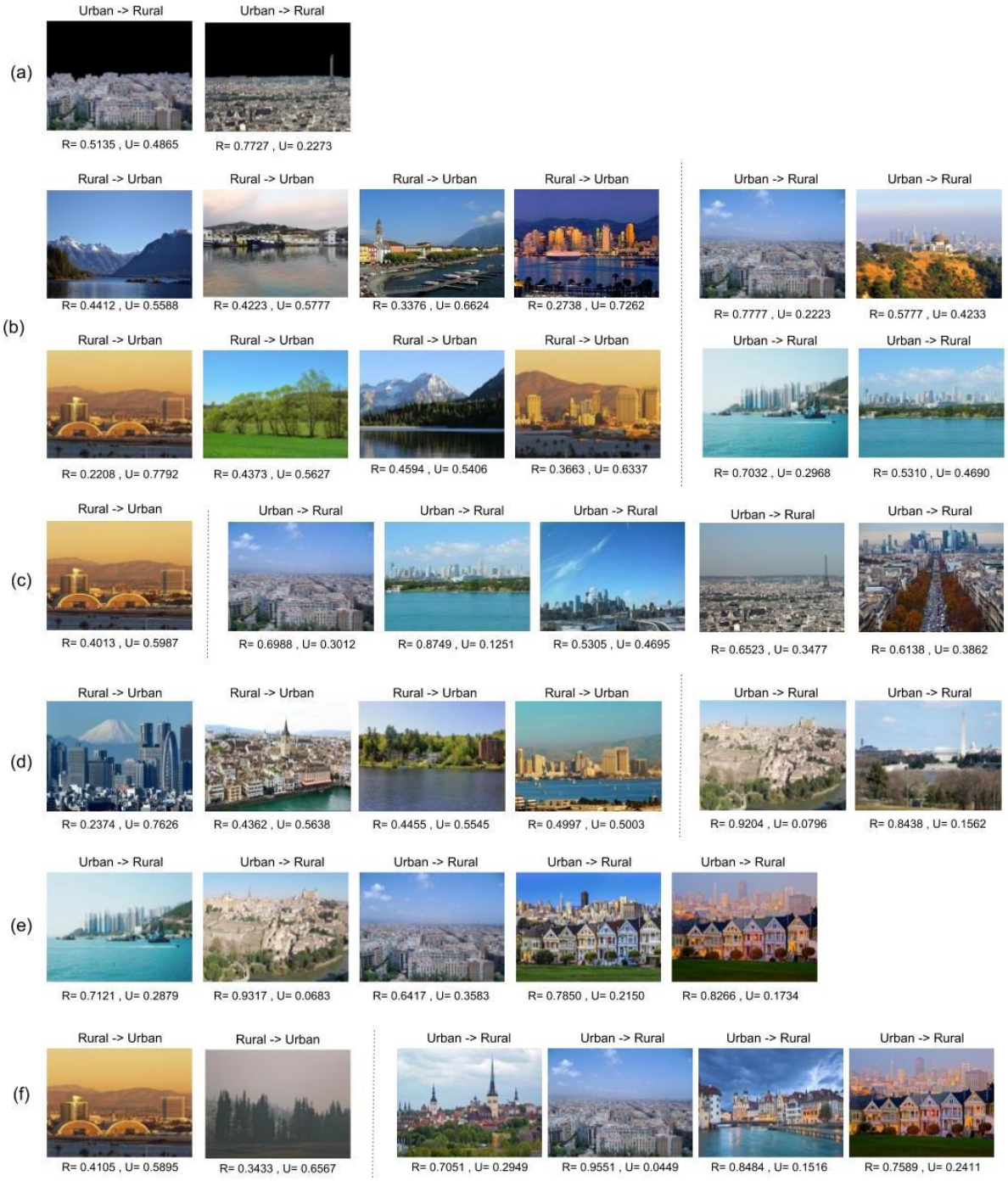

Fig.8. Miss Classified Skylines with the obtained accuracies; U: Urban, R: Rural (a) Sky-CNN (b) Inception-ResNet-v2 [31] (c) Inception-v3 [31] (d) ResNet-50 [33] (e) ResNet-101 [25] (f) GoogLeNet [9]

Copyright © 2019 MECS

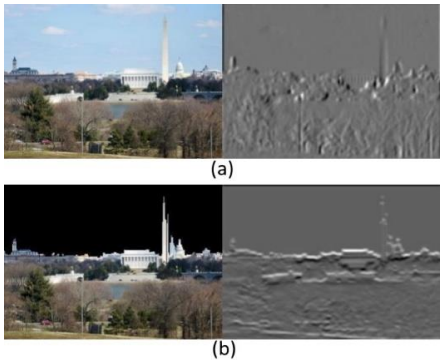

Fig.9. The Strongest Activations of the first convolutional layer for the Washington Skyline using (a) fine-tuned ResNet-50 (b) Sky-CNN

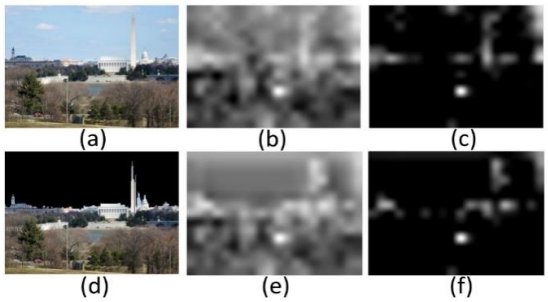

Fig.10. The third Activation channels of (b)(e) the 26th convolutional layer and (c)(f) the 26thReLU layer for the Washington Skyline

VI. Conclusions

This paper introduces a new CNN-based approach for skyline scenes context conceptualization taking under consideration the pertinent information worn by the under horizon-line part in skyline scenes. The appropriate residual deep network model adopted and customized to categorize skyline scenes to urban or rural was an effective line that outperforms hand-engineered approaches describing such scenes.

The proposed model Sky-CNN succeeds in establishing a deep representation of Skyline scenes by combining the appropriate features and highlighting the deeply-informative under-skyline area. The pre-processing step considering the horizon line that distinguishes the landscapes images was effective to increase the classification accuracies of urban and rural Skylines.

Acknowledgments

This work was funded by the “ANR-12-VBDU-0008 - Skyline” project of the “Agence Nationale de la Recherche (ANR)”, and by the the LabEx “Intelligence des mondes Urbains - IMU”.

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

[1] Wei, X., Phung, S.L., Bouzerdoum, A.: ‘Visual descriptors for scene categorization: experimental evaluation’, Artificial Intelligence Review, 2016, 45, (3), pp. 333–368. Available from: https://doi.org/10.1007/s10462-015-9448-4

[2] Sassi, A., Amar, C.B., Miguet, S. ‘Skyline-based approach for natural scene identification’. In: 13th IEEE/ACS International Conference of Computer Systems and Applications, AICCSA 2016, Agadir, Morocco, November 29 - December 2, 2016. pp. 1–8.

[3] Day, A.: ‘Urban visualization and public inquiries: the case of the heron tower, london’, Architectural Research Quarterly, 2002, 6, (4), pp. 363–372

[4] III, A.S., Nasar, J.L., Hanyu, K.: ‘Using pre-construction validation to regulate urban skylines’, Journal of the American Planning Association, 2005, 71, (1), pp. 73–91

[5] Nasar, J.L., Terzano, K.: ‘The desirability of views of city skylines after dark’, Journal of Environmental. Psychology, 2010, 30, (2), pp. 215 – 225

[6] Ayadi, M., Suta, L., Scuturici, M., Miguet, S., Ben.Amar, C. In: Blanc.Talon, J., Distante, C., Philips, W., Popescu, D., Scheunders, P., editors. ‘A parametric algorithm for skyline extraction’. (Cham: Springer International Publishing, 2016. pp. 604–615

[7] Tonge, R., Maji, S., Jawahar, C.V. ‘Parsing world’s skylines using shape-constrained mrfs’. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 3174–3181

[8] Sassi, A., Ouarda, W., Ben.Amar, C., Miguet, S. ‘Neural Approach for Context Scene Image Classification based on Geometric, Texture and Color Information’. In: Representation, analysis and recognition of shape and motion FroM Image data. (Aussois, France: RFIA, 2017. Availablefrom: https://hal.archives-ouvertes.fr/hal-01687973

[9] Yassin, F.M., Lazzez, O., Ouarda, W., Alimi, A.M. ‘Travel user interest discovery from visual shared data in social networks’. In: 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT), pp. 1–7

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. ‘Going deeper with convolutions’. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. pp. 1–9

[11] Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., et al. ‘Convolutional recurrent neural networks: Learning spatial dependencies for image representation’. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015. pp. 18–26

[12] Gong, Y., Wang, L., Guo, R., Lazebnik, S.: ‘Multi-scale orderless pooling of deep convolutional activation features’, CoRR, 2014, abs/1403.1840. Available from: http://arxiv.org/abs/1403.1840

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

[14] Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: ‘Image classification with the fisher vector: Theory and practice’, Int J Comput Vision, 2013, 105, (3), pp. 222–245

Copyright © 2019 MECS

[15] Yang, J., Yu, K., Gong, Y., Huang, T.S. 'Linear spatial pyramid matching using sparse coding for image classification'. In: CVPR. (IEEE Computer Society, 2009. pp. 1794–1801
[16] Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: 'Sun database: Exploring a large collection of scene categories', International Journal of Computer Vision, 2016, 119, (1), pp. 3–22
[17] Oliva, A. & Torralba, A.: 'Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope'. In: International Journal of Computer Vision, 2001, 42: 145–147.
[18] Ojala, T., PietikÃd'inen, M., Harwood, D.: 'A comparative study of texture measures with classification based on featured distributions', Pattern Recognition, 1996, 29, (1), pp. 51 – 59
[19] Huttunen, S., Rahtu, E., Kunttu, I., Gren, J., Heikkilä, J. In: Heyden, A., Kahl, F., editors. 'Real-time detection of landscape scenes'. (Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. pp. 338–347
[20] Han, X., Chen, Y. 'Image categorization by learned PCA subspace of combined visual-words and low-level features'. In: Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2009), Kyoto, Japan, 12-14 September, 2009, Proceedings, 2009. pp. 1282–1285
[21] Serrano, N., Savakis, A.E., Luo, J.: 'Improved scene classification using efficient low-level features and semantic cues', Pattern Recognition, 2004, 37, (9), pp. 1773– 1784
[22] Vailaya, A., Jain, A., Zhang, H.J.: 'On image classification: City images vs. landscapes', Pattern Recognition, 1998, 31, (12), pp. 1921 – 1935
[23] Chen, Z., Chi, Z., Fu, H. 'A hybrid holistic/semantic approach for scene classification'. In: 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014. (, 2014. pp. 2299–2304
[24] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P. 'Gradient-based learning applied to document recognition'. In: Proceedings of the IEEE. (, 1998. pp. 2278–2324
[25] He, K., Zhang, X., Ren, S., Sun, J. 'Deep residual learning for image recognition'. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. pp. 770–778
[26] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. 'Tensorflow: A system for large-scale machine learning'. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. OSDI'16. (Berkeley, CA, USA: USENIX Association, 2016. pp. 265–283. Available from: http://dl.acm.org/citation.cfm?id=3026877.3026899
[27] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., et al.: 'A survey on deep learning in medical image analysis', Medical Image Analysis, 2017, 42, pp. 60 – 88. Available from: http://www.sciencedirect.com/science/article/pii/S1361841517301135
[28] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 'Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40, (4), pp. 834–848
[29] Balduzzi, D., Frean, M., Leary, L., Lewis, J.P., Ma, K.W., McWilliams, B.: 'The shattered gradients problem: If resnets are the answer, then what is the question?', CoRR, 2017, Available from: http://arxiv.org/abs/1702.08591
[30] Philipp, G., Song, D., Carbonell, J.G.. 'Gradients explode - deep networks are shallow - resnet explained', 2018. Available from: https://openreview.net/forum?id=HkpYwMZRb
[31] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: 'Rethinking the inception architecture for computer vision', CoRR, 2015, abs/1512.00567. Available from: http://arxiv.org/abs/1512.00567
[32] He, K., Zhang, X., Ren, S., Sun, J.: 'Identity mappings in deep residual networks', CoRR, 2016, abs/1603.05027. Available from: http://arxiv.org/abs/1603.05027
[33] Hiippala, T. 'Recognizing military vehicles in social media images using deep learning'. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), 2017. pp. 60–65
[34] Alvarez, S., Vanrell, M.: 'Texton theory revisited: A bag-of-words approach to combine textons', Pattern Recognition, 2012, 45, (12), pp. 4312– 4325.

Authors' Profiles

Ameni Sassi is a Computer Engineer since 2011. She received theMastersdegree in New Technologies and Domain-Specific Computer Systems from theNational School of Engineers of Sfax, University of Sfax, in 2013. Currently, she is a Ph.D student in Computer Systems Engineering.

Wael Ouarda received a Master Degree in Computer Science: Knowledge and Decision from the INSA Lyon in France in 2010. He is now a PhD in Research groups on Intelligent Machines from the National School of Engineers of Sfax. His current research interests include Soft Biometrics, Information Fusion, SOA Approach for IT and Optimization Patterns.

Chokri Ben Amar received the B.S. degree in Electrical Engineering from the National Engineering School of Sfax (ENIS) in 1989, the M.S. and PhD degrees in Computer Engineering from the National Institute of Applied Sciences in Lyon, France, in 1990 and 1994, respectively. He spent one year at the University of "Haute Savoie" (France) as a teaching assistant and researcher before joining the higher School of Sciences and Techniques of Tunis (ESSTT) as Assistant Professor in 1995. In 1999, he joined the Sfax University (USS), where he is currently a professor in the Department of Electrical Engineering of the National Engineering School of Sfax. His research interests include ComputerVision and Image and video analysis. These research activities are centered on Wavelets and Wavelet networks and their applications to data Classification and approximation, Pattern Recognition and image and video coding, indexing and watermarking.

Copyright © 2019 MECS

Serge Miguet graduated from the ENSIMAG (Grenoble, France) in 1988. He obtained a PhD from the INPG in 1990. He was an Assistant Professor at the ENS de Lyon, and a member of the LIP laboratory from 1991 to 1996. He received his Habilitation Diriger des Recherches from the Université Claude Bernard Lyon 1 in 1995. Since 1996, he is a full Professor in Computer Science at the Université Lumière Lyon 2, and a member of the LIRIS laboratory, UMR CNRS 5205. His main research activities are devoted to models and tools for image processing, image analysis, shape recognition.

How to cite this paper: Ameni Sassi, Wael Ouarda, Chokri Ben Amar, Serge Miguet, "Sky-CNN: A CNN-based Learning Approach for Skyline Scene Understanding", International Journal of Intelligent Systems and Applications(IJISA), Vol.11, No.4, pp.14-25, 2019. DOI: 10.5815/ijisa.2019.04.02

Copyright © 2019 MECS