Neural Approach for Context Scene Image Classification based on Geometric, Texture and Color Information

Ameni Sassi[1], Wael Ouarda[1], Chokri Ben Amar[1] and Serge Miguet[2]
[1] *REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax, ENIS, BP 1173, 3038, Sfax, Tunisia*
[2] *LIRIS, Université de Lyon, UMR CNRS 5202, Université Lumiére Lyon 2, 5 av. Mendès-France, Bât C, N 123, 69676. Bron, Lyon ,France*

{*ameni.sessi.tn,wael.ouarda,chokri.benamar*}*@ieee.org, serge.miguet@univ-lyon2.fr*

Abstract. Revealing the context of a scene from low-level features representation, is a challenging task for quite a long time. The classification of landscapes scenes to urban and rural categories is a preliminary task for landscapes scenes understanding. Having a global idea about the scene context (rural or urban) before investigating its details, would be an interesting way to predict the content of that scene. In this paper, we propose a novel features representation based on skyline, colour and texture, transformed by a sparse coding using Stacked Auto-Encoder. To evaluate our proposed approach; we construct a new database called SKYLINEScene Database containing 2000 images of rural and urban landscapes with a high degree of diversity. Many experiments were carried out using this database. Our approach shows it robustness in landscapes scenes classification.

keywords: Deep Neural Network; Auto-Encoder; Scene Classification; Skyline and Curvature

1 Introduction

Having a wider idea about the preference of people to the skyline of the cities where they live or they want to visit , is an important issue in social urbanism. This study was in the framework of a sustainable city project SKYLINE. The main purposes of that project is the identification and the systematic-analyze of the landscapes perceptions of the general public and practitioners by corresponding the aspects taken from the skyline photographs and the perceptions collected from an interesting number of audiences within European cities (The example of London and Lyon). One of our distant goals is to objectify the effect of natural elements such as vegetation and mountains on the representations of urban landscapes using a photo-questionnaire system.

The first step to achieve our goals is to reveal, from a landscape photo, if it represents a city or a rural scene, by scanning the whole skyline. The Skyline could be defined as the silhouette describing a place, or in other words, the profile of some cities or towns or different places. The nature of a skyline is an important cue on evaluating the landscapes perceptions. So, classifying a landscape scenes into urban and rural ones would be a sufficient first step for our purposes. This study is dedicated to the classification of landscapes scenes based on a deep neural approach with a new combination of some features which are geometric, texture and color.

The rest of the paper is composed of four major sections. The first one exposes some related works. Then, in the second part, we will describe our proposed neural approach for landscapes classification. After that we present our constructed database and some first results evaluating our proposed approach. The last section contains the conclusions about the realized works and some perspectives that will be the goal of a future work.

2 Related works

Natural scenes classification is an interesting task for a variety of applications of computer vision (content-based image retrieval systems [5], pattern recognition, image understanding). This topic can touch many facets of computer vision like scene segmentation or labelling, scene parsing or object detection [4].

The work [3] proposed a hybrid holistic/semantic approach for natural scenes classification. Using the Hierarchical Matching Pursuit (HMP) to learn holistic features and the Semantic Spatial Pyramid SSP to represent the spatial object information, this work combined these two strategies with a support vector machine (SVM) 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 touching the facet of scene parsing [13] proposed an approach for outdoor scenes classification. The first step in their process is to generate Spatially Constrained Location Prior (SCLP). The second one is the prediction of class probabilities using visual feature based classifiers. This last step is followed by the propagation of contextual class votes based on SCLP to reach the final step which is the integration of visual feature based class probabilities and contextual class votes. The visual features used in that work were the RGB histograms in addition to the texton and SIFT histograms. They adopted the SVM with radial basis function kernel as a classifier. The performance of the proposed approach in [13] was evaluated in two datasets (The Stanford Background and SIFT Flow) and gave a global accuracy of 81.2% and a class accuracy of 71.8% for the first dataset. Reviewing the approaches that combine object detection and scene classification, we found out that work [12] that proposes CRF (conditional random field) models reasoning jointly about object detection, image labelling and scene classification. To create the unitary potentials (composing their holistic model) for the scenes, they used a standard bag-of-words spatial pyramid with a sparse coding dictionary on RGB histograms, color moment invariants, SIFT features and colorSIFT, and trained a linear one-vs-all SVM classifier. The scene classification accuracy reached in this work was 80.6% on the MSRC-21 dataset (origMSRC).

3 Proposed system

As shown in the Fig. 1 the proposed system is composed of three main processing steps: 1) the features extraction from landscapes images based on the skyline and the texton, 2) the training of a deep neural network (sparse autoencoders), and 3) the classification process based on the extracted visual features.
We used a deep neural network to achieve the task of skylines classification based on the geometric and the texture-color features. Thus, we train a neural network using two hidden layers. These hidden layers are trained individually using autoencoders. After that, for the classification process, we compared the results using a support vector machine and a softmax classifier used is the Support Vector Machines.

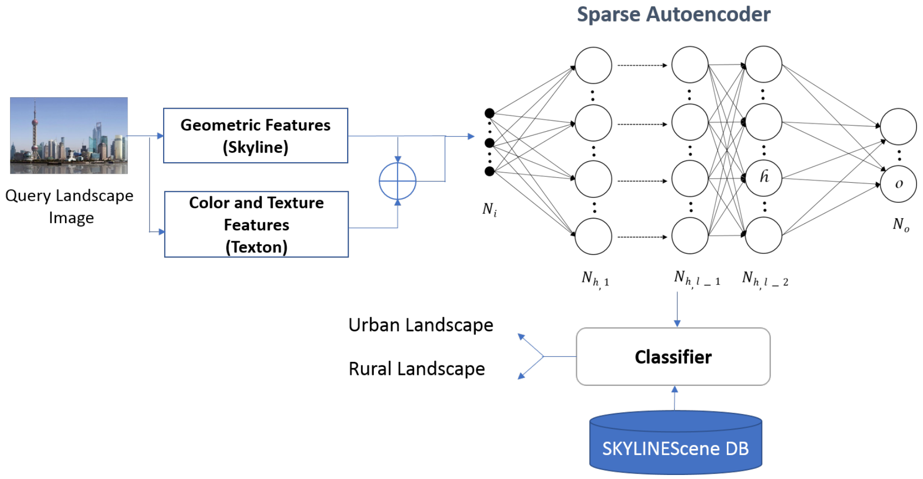

Fig. 1: Architecture of the proposed system

3.1 Features Extraction

3.1.1 Skylines Geometric Description

Based on the sky line extracted from landscapes scenes, we took out suitable measurements allowing to distinguish between rural landscapes and urban landscapes.

The straight lines classification The idea was to take a look at the skyline as a polyline and determine if there is an important number of straight lines. So, we begin by extracting the straight line segments based on the Douglas-Peucker approximation algorithm. Then, we compute the resulting segments length. After that, we create categories based on segments length distribution; and then counting segments by length category to get a histogram presenting the number of segments on each category of segments length. The limits of categories intervals follow the form of geometric sequence and it depend on three main elements which are: (i) the shortest segment length within the approximated skyline, (ii) the number of the categories itself, and (iii) the height of the landscape image. The used sequence helps in having an appropriate histogram for each landscape image and avoids getting categories with no segments. An example of the obtained histograms is shown in Fig. 2 (c).

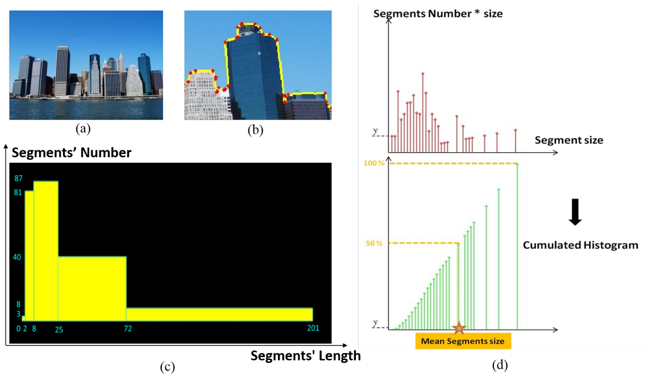

Fig. 2: (a) The cityscape of New York (b) The polygonal approximated segments of the skyline (c) The segments classification histogram (d) Modified cumulated histogram

To characterize each skyline with concrete values, we generate histograms with different distribution in order to highlight the linkage between the number of segments per skyline and their length and get a significant value of the mean segments size (Fig. 2 (d)). The mean segments size, derived from these cumulated histograms, and the length of the longest segment on the skyline could be an indicative couple of values that describes the skyline in a global way: the skyline is almost artificial or natural. The obtained values and more details are illustrated on [11]. These values are illustrated in the second and third column of the table 1 for some samples of skylines.

Table 1: Geometric descriptive values for some skylines

Landscape Image	**Index of the Mean Segments size**	**Longest segment**	**% Percentage of resisting key points at the middle scale**
New York	24	68	14.444444
Dubai	40	114	10.465116
London	24	50	8.928571
Rural Landscape 1	18	41	3.773585
Rural Landscape 2	16	23	6.703911
Rural Landscape 3	19	26	4.83871

The curvature analysis By examining the skyline curve, we can affirm that the curvature goes through a lot of changes. It is clear that the mountain peaks, the indentation created by a vegetation, and the structure of a building have very different values of curvature. To calculate the curvature, we admit the formula (4) where the skyline is the curve defined by (5).

$$k(t) = \frac{\dot{x}(t)\ddot{y}(t) - \ddot{x}(t)\dot{y}(t)}{\left(\dot{x}^2(t) + \dot{y}^2(t)\right)^{3/2}} \tag{1}$$

$$\boldsymbol{c} = (x(t), y(t)) \tag{2}$$

To well determine the geometric features along the skyline and select the most important key points, we used the curvature scale-space description CSS. This descriptor was at the first time introduced by [7] and used for a variety of applications such as : the shape similarity retrieval [6] and the corner detection [8]. The process comprises the computation of the curvature values for a curve that has to be smoothed progressively on each scale via Gaussian kernels. So, we have a different set of curvatures values at each smoothing scale. The skylines, we get in different smoothing scales for different rural landscapes images, depict an important number of key points that replicate the fast variation in the curvature, nevertheless, these points vanished from low smoothing scales. In contrast, for urban landscapes, the corners of buildings or any remarkable point fight till high scales of smoothing. To interpret these notes with values, we represented the waning of the key points number across different scales of smoothing as scatter plots (Fig. 3). The concrete number describing these graphics we picked, is the percentage of key points that resist until the middle smoothing scale . To validate our observations, these percentages for cityscapes should be higher than the ones we got for rural landscapes [11]. Some percentages for diverse landscapes are shown in table 1. These plotted curves validate the short lifetime of the key points across the scales in rural landscapes, unlike, the case for urban landscapes where these points persist until very high scales.

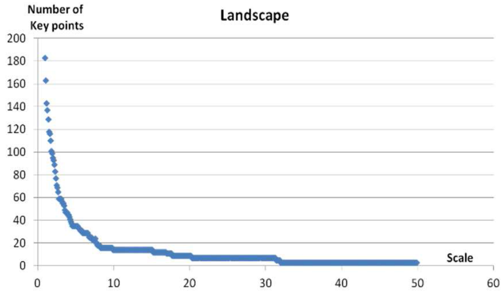

Fig. 3: The lifetime of skyline key points over smoothing scales

3.1.2 Colour and Texton for Skylines

Colour and texture are frequently used as low-level features for image classification. Looking at the part of the image under the skyline, we can notice that the colour and the texture presenting buildings are different from the ones describing mountains or vegetation. Then, the colour and the texture could be discriminative features to distinguish between landscapes with urban skyline and rural landscapes. Searching on recent works proposing a combination between colour and texture features, we found this one [1] that suggests to compute the image textons following the original definition [2] and adding the color information. Starting with the definition of texton as blob attributes, this work proposed a texture representation based on Bag of Words framework, that represents the texture-colour image content.

For the colour-texture representation of our landscapes images, we applied the co-joint texton descriptor (JTD). This descriptor is defined as the probability density function of a bidimensional random variable(C,S); C concerns the quantised colour texton space and S concerns the quantised shape texton space. The Fig. 4 depicts an example of a JTD descriptor. This colour texture descriptor were applied to our SKYLINEScene database to represent the part under skyline and not the whole image, as shown in the Fig. 5.

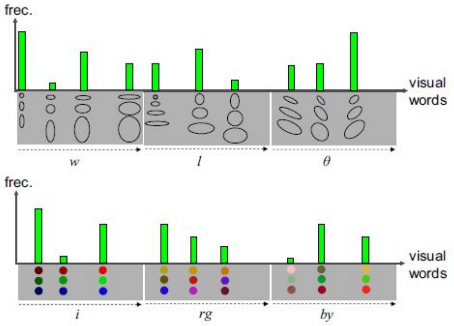

Fig. 4: A schematic representation of the JTD

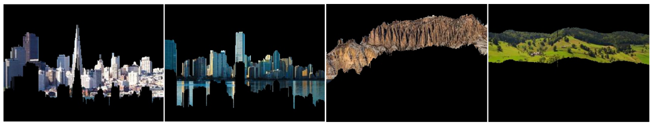

Fig. 5: The Under-skylines images

3.2 Deep Neural Network Training

The deep architecture used in our system was a stacked auto-encoders architecture that has a series of inputs, outputs and hidden layers. So for that, we have used the nonlinear auto-encoders (Fig. 6) to construct each hidden layer of the deep neural network. The input layer of the first hidden layer (first auto-encoder) is the input layer for the all network. Starting from the second hidden layer (Second autoencoder), there is always reconstructing of the output of the previous layer. Namely, for each layer, we have reconstructed features using a number of neuron smaller than the number of neuron of the previous one/The number of neuron for the first auto-encoder is bigger than the number of features. Using this method, we have constructed a deep neural network with two hidden layers using the extracted geometric skyline features. We have trained the two hidden layers individually using two unsupervised autoencoders. The objective of the reconstruction of features and reducing the number of neuron of each hidden layer is to force the network to learn only the most important features and achieve a dimensionality reduction and separate the maximum between features of classes [9][10]. Finally, we obtained an unsupervised neural network which is shown in the Fig. 6.

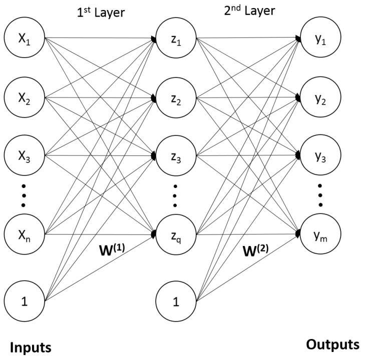

Fig. 6: Unsupervised Neural Network Architecture

3.3 Classification

3.3.1 The SVM Classifier

The simple Support Vector Machines Classifier performs with low complexity than other kernels such as sigmoid, radial basis function, polynomial. For the SVM classification the challenge is to find the appropriate hyper plane that separates the data in two classes (positive and negatives). The used architecture of multiSVM aims to construct two groups separated by the optimal hyper plane. We can find more than one hyper plane. In fact, another problem appear when the data are not in linear possibility of separation. The strength of SVM classifier compared to the Neural Network method is that the SVM is capable to overcome with the convergence problem in a local minimum of the optimization function. It scales relatively well to low dimensional data and the trade-off between classifier complexity and error can be explicitly controlled.

3.3.2 The Softmax Classifier

The SoftMax classifier denoted as SMC is a supervised model which generalizes logistic regression as

$$f_{w^{(3)}}(z) = \frac{1}{1 + exp(-W^{(}3)^{T} z)} \tag{3}$$

where $f_{w^{(3)}}$ is a sigmoid function having as parameters $W^{(3)}$. When the input z of the softmax classifier is a high-level representation of the skyline features learned by the Stacked Sparse Auto-Encoder, the Softmaxs parameter $W^{(3)}$ is trained with the set explained in (...) to minimize the cost function. By minimizing the cost function with respect to $W^{(3)}$ via the gradient descent based approach, the parameter $W^{(3)}$ can then be determined.

$$\left\{h^{(2)}(k), y(k)\right\}_{k=1}^{N} \tag{4}$$

4 Experiments and Results

In this section, we evaluate our proposed approach on SKYLINEScene database. The photographs taken into consideration in our constructed database are the ones showing the global view of cities or rural places.We experiment, first, the use of the geometric features extracted from the sky Line to get the classification accuracies. To find the Auto-Encoder neural network architecture that gives the best accuracies results, we have experimented three architectures with a final SoftMax layer. The results are summarized in Table 2. To reveal if the SVM classifier gives better results than the SoftMax classifier, we make some tests using different SVM kernel functions. The experimental configuration for the SVM classifier is as follow: the dataset is randomly divided into ten folds; one fold for the test and the lasting nine folds for training. The average performance of ten folds testing data is reported. The parameters of SVMs are set by two-fold cross-validation on the training data. The reported overall performance is the average accuracy of the two classes. Table 3 illustrates the accuracies and the standard deviation values for three different kernel functions.

Table 2: Classification accuracies using a softmax classifier

Architecture		**Accuracies**	**Urban Accuracy**	**Rural Accuracy**
1st HL	2nd HL			
300	150	**85.78%**	88.12%	83.7%
30	15	85.7%	88.44%	83.3%
20	10	85.68%	88.34%	83.32%

Table 4 shows the classification accuracy obtained using our proposed neural approach based on the combination of the geometric, the color and the texture features from Skylines. These accuracies show up the usability of the geometric

Table 3: Classification accuracies using an SVM classifier

kernel Function	Accuracy	Standard Deviation
Linear	82.86%	0.006
RBF (sigma=5)	**83.87%**	0.007
Polynomial (3 Planes)	78.09%	0.0381

features extracted from skylines in the classification of landscapes scene since this horizon line is a very specific feature to landscapes images. Combining the geometric features with the texton ones, the classification accuracy reaches 84.92%.

Table 4: Classification accuracies depending on features

Features Vector	Accuracy	Standard Deviation
Geometric (Skyline)	83.87%	0.007
Color and Texture (Texton)	62.97%	0.0012
Geometric+Color+Texture	**84.92%**	0.001

To have a global vision about the performance of our proposed system for context scene classification, we summarized in the Table 5 the related works mentioned in the state of the art. We can notice that there is obviously differences between the number of classes and also the images per classes. Our approach proves it robustness in landscapes scenes classification behind the existing approaches.

Table 5: Comparison of our system with the state of the art related works

Works	Datasets	Image per Class	Scene Classes	Total images	Image size	Results
[12]	MSRC-21	∼13	21	591	320x213	80.6%
[3]	Natural scene dataset	∼100	6	700	480x720	78.2%
[13]	The Stanford background	∼90	8	715	320x240	81.2%
Our approach	SkylineScene	1000	2	2000	320x240	84.92%

5 Conclusions

This paper introduce a new neural approach for landscapes scenes classification based on a very specific feature which is the skyline and the color-texture features. To represent these combination of low-level features, we build a sparse stacked Auto-Encoders architecture to have a new structure of our input data. The new SKYLINEScene database, containing a specific collection of rural and urban landscapes photographs, was created to evaluate our approach. The classification accuracies reached are very competitive and they confirm that the skyline is a significant geometric feature for landscapes scenes.

Our work with the geometric features of the skyline should be expanded by using other tools to describe better the skyline. The results obtained using our skyline-based approach to classify landscapes scenes will be compared with similar works based on a variety of local and global features.

Acknowledgements 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.
This work was initiated in the project SKYLINE which is led by EVS (Lyon2), with the following participants: lEIVP-Ville de Paris, le LIRIS (Lyon1-2) and ”lAgence dUrbanisme de Lyon.”

References

1. Susana Alvarez and Maria Vanrell. Texton theory revisited: A bag-of-words approach to combine textons. *Pattern Recognition*, 45(12):4312 – 4325, 2012.
2. James R. Bergen and Béla Julesz. Rapid discrimination of visual patterns. *IEEE Trans. Systems, Man, and Cybernetics*, 13(5):857–863, 1983.
3. Zenghai Chen, Zheru Chi, and Hong Fu. A hybrid holistic/semantic approach for scene classification. In *22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014*, pages 2299–2304, 2014.
4. Onsa Lazzez, Wael Ouarda, and Adel M. Alimi. *Age, Gender, Race and Smile Prediction Based on Social Textual and Visual Data Analyzing*, pages 206–215. Springer International Publishing, Cham, 2017.
5. Onsa Lazzez, Wael Ouarda, and Adel M. Alimi. *Understand Me if You Can! Global Soft Biometrics Recognition from Social Visual Data*, pages 527–538. Springer International Publishing, Cham, 2017.
6. Farzin Mokhtarian and Sadegh Abbasi. Shape similarity retrieval under affine transforms. *Pattern Recognition*, 35(1):31–41, 2002.
7. Farzin Mokhtarian and Alan K. Mackworth. Scale-based description and recognition of planar curves and two-dimensional shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(1):34–43, 1986.
8. Farzin Mokhtarian and Riku Suomela. Robust image corner detection through curvature scale space. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1376–1381, 1998.
9. Hanen Nasri, Wael Ouarda, and Adel M. Alimi. Relidss: Novel lie detection system from speech signal. *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8, 2016.
10. Wael Ouarda, Hanene Trichili, Adel M Alimi, and Basel Solaiman. Towards a novel biometric system for smart riding club. *Journal of Information Assurance & Security*, 11(4), 2016.
11. Ameni Sassi, Chokri Ben Amar, and Serge Miguet. 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*, pages 1–8, 2016.
12. Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702–709. IEEE Computer Society, 2012.
13. Ligang Zhang, Brijesh Verma, David R. B. Stockwell, and Sujan Chowdhury. Spatially constrained location prior for scene parsing. In *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*, pages 1480–1486, 2016.