Query-by-example HDR image retrieval based on CNN

Raoua Khwildi, Azza Ouled Zaid, Frédéric Dufaux

Noname manuscript No.
(will be inserted by the editor)

Query-by-example HDR image retrieval based on CNN

Raoua Khwildi · Azza Ouled Zaid · Frédéric Dufaux

Received: date / Accepted: date

Abstract Due to the expension of High Dynamic Range (HDR) imaging applications into various aspects of daily life, an efficient retrieval system, tailored to this type of data, has become a pressing challenge. In this paper, the reliability of Convolutional Neural Networks (CNN) descriptor and its investigation for HDR image retrieval are studied. The main idea consists in exploring the use of CNN to determine HDR image descriptor. Specifically, a Perceptually Uniform (PU) encoding is initially applied to the HDR content to map the luminance values in a perceptually uniform scale. Afterward, the CNN features, using Fully Connected (FC) layer activation, are extracted and classified by applying the Support Vector Machines (SVM) algorithm. Experimental evaluation demonstrates that the CNN descriptor, using the VGG19 network, achieves satisfactory results for describing HDR images on public available datasets such as PascalVoc2007, Cifar-10 and Wang. The experimental results also show that the features, after a PU processing, are more descriptive than those directly extracted from HDR contents. Finally, we show the superior performance of the proposed method against a recent state-of-the-art technique.

Keywords Image retrieval · Convolutional Neural Networks · High Dynamic Range · Feature extraction · VGGNet · Perceptually Uniform encoding

R. Khwildi and A. Ouled Zaid
Laboratoire Systèmes de Communications
École Nationale d'Ingénieurs de Tunis, Université de Tunis El Manar
B.P. 37 le Belvédère 1002 Tunis, Tunisie
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: khwildi.raoua135@gmail.com, azza.ouledzaid@isi.rnu.tn

F. Dufaux
Laboratoire des Signaux et Systèmes
Université Paris-Saclay, CNRS, CentraleSupelec
91190 Gif-sur-Yvette, France
E-mail: frederic.dufaux@l2s.centralesupelec.fr

1 Introduction

HDR imaging has received a lot of attention in modern computer graphic applications. Its success is mainly due to its ability to capture an extremely wide range of the illumination in real-world scenes and to produce images that are more realistic. Numerically, the HDR image is encoded by three floating-point numbers related to the physical luminance in the scene; typically with 96 bit per pixel (bpp) instead of 24 bpp in its Low Dynamic Range (LDR) version.

Over the past decades, HDR imaging [1] has received significant recognition in several computer vision tasks [2,3]. As a result, the subject of HDR image acquisition [4,5] has attracted the attention of researchers and raised the challenge of storing the generated HDR images in specific formats. In this context, various compression methods have been proposed to represent the floating-point numbers in an efficient and compact way. Several formats support these types of data like RGBE [6], LogLuv [7], and OpenEXR [8] formats. However, the use of these formats is hampered by difficulties in rendering the HDR content on standard display devices that are designed for conventional images. This problem has been tackled by using tone mapping operators [9], which aim at reducing the high dynamic range while preserving the image content such as contrast, brightness, and colors. On the other hand, some researchers [10,11,12,13,14,15,16] are interested in developing reverse tone mapping (rTM) methods to expand LDR content to HDR. The principle of rTM is to estimate, from the LDR image, the real-world luminance values as faithfully as possible.

In accordance with the development of HDR imaging, it is expected that the number of HDR images will grow rapidly and that collections of this type of images will become available in different application domains. Therefore, the development of effective HDR image indexing and retrieval methods is becoming extremely important. In the literature, many works have focused on LDR image retrieval using different methods. In the last few years, deep learning approaches have become foremost choice to address most problems in the fields of computer vision and image processing like Image Dehazing [17,18], Recommender System [19], Object Detection [20], Visual Captioning[21] and Image retrieval [22,23,24,25,26,27]. The latter is a fundamental task in many computer vision applications. It gained the interest of the scientific community to access, search or browse effectively the images from databases. Several CNN based methods have been developed in this field to supply a high-level description of image content. In section 2, we introduce some of them.

The CNN architecture provides an attractive solution for different tasks thanks to its high performance, discriminative power, and compact representation, allowing for a large-scale data modeling. However, to the best of our knowledge, no CNN-based scheme has been proposed yet for the purpose of HDR image retrieval. In this paper we aim to shed light on this issue. Specifically, we propose a query-by-example HDR image retrieval method that uses the Fully Connected (FC) layer activation to define the relevant features of

HDR images. Before passing through the descriptor computation stage, the HDR pixels are modified by using a Perceptually Uniform (PU) encoding [28] to map the luminance values in a perceptually uniform scale. The originality of our approach lies in extracting and testing the CNN features on the HDR contents. To this end, we selected the method that has powerful descriptors and high HDR retrieval accuracy. The novel contributions of this work are listed as follows:

- Design an algorithm for HDR image retrieval based on CNN.
- Apply PU encoding [28] to HDR content and evaluate its influence on the retrieval accuracy.
- Build an HDR image database for the purpose of retrieval performance evaluation.
- Analyse the efficiency of CNN descriptor and report the performance of Visual Geometry Group Network (VGGNet).
- Evaluate the effectiveness of the proposed retrieval algorithm according to the number of layers in CNN.
- Demonstrate the competitiveness of conventional and FC layers for LDR and HDR datasets.
- Present experiments showing significant accuracy improvements on recent state-of-the-art method.

The paper is structured as follows: In Section 2, we give a brief overview on the related works regarding HDR image retrieval and the use of CNN methodology. In Section 3 we present the commonly used CNN architecture. In Section 4, we describe the proposed CNN-based scheme for HDR image retrieval. In the experimental section 5 we compare our method to other ones in the literature and assess their accuracy. Finally, conclusions are drawn in Section 6.

2 Related work

In the last few years, some works have focused on HDR image retrieval. In [29], the authors proposed to use histogram intersection to define an HSV color descriptor. The results of the experiments have revealed that HSV histograms can be efficiently used as a global descriptor for HDR image retrieval task. To ameliorate their method, the authors in [30,31] combined the HSV color histograms with color moments. Despite their practical use, these approaches [30,31,29] seem to be very limited compared to the abilities of local descriptors that have been proved to be very effective for indexing applications. Some researchers [32,33,34] turned their attention to the detection of key-points in HDR images, under changing illumination conditions, varying camera viewpoints, camera distances and scene lighting. Experimental results, reported in [32,34], demonstrated that the direct use of HDR image in a linear scale is inappropriate for key-points detection. In [35], the authors introduced a new retrieval method based on LDR expansion. They improved the feature extraction by using reverse tone mapping and applying a tone mapping operator

to determine the Scale Invariant Feature Transform (SIFT) descriptor. The experimental results showed the potential of the tone-mapped HDR content for detecting the local descriptors and demonstrated that the selected features are more descriptive than those extracted from LDR and HDR versions. The authors in [35] also established that the use of local SIFT descriptors is not appropriate for HDR images.

Recently, to achieve a higher level of robustness, researchers have successfully used machine learning approaches in many imaging applications. Generally speaking, deep learning systems allow building rich features with hierarchical representation, resulting is an effective classification [40]. Particularly, CNN architecture becomes one of the most interesting topics that revolutionized the field of computer vision like segmentation and object detection [36,37]. It is characterized by its ability to capture different patterns while achieving a high classification accuracy. In literature, several systems, based on CNN, have been investigated to effectively describe LDR images. Among these, some methods use the activations of fully connected or convolutional layers as image descriptors [23,24,25,26,27]. In [37], authors propose off-the-shelf CNN features. They extract generic features from OverFeat network using the fully connected (FC6) of AlexNet and demonstrate that this approach clearly outperforms local features methods. Various works use the activations of max-pooling from convolutional layers like [54]. To obtain compact descriptors, a number of dimensionality reduction methods are applied like Principal Component Analysis (PCA) [25,27], Bag-of-Words [55], VLAD [42] and Fisher Vectors [56]. Authors in [57] propose to use a trainable Generalized-Mean (GeM) pooling layer. The idea consists in adding a new pooling layer with learnable parameters after the convolutional layers. Then, a whitening is applied to reduce the descriptor dimensionality. In [57], the authors introduce a new weighted query expansion. The work presented in [58] consists in building a descriptor based on the regional maximum activations of convolutions (R-MAC) descriptor [27] and learn CNN weights in an end-to-end method, and applying the siamese network with three streams and a triplet loss for training. Regional network is proposed to select the relevant regions of the image, using image scaling to extract local features. Other solution consists in adding a new layer named NetVLAD that can be applied in any CNN architecture. It is trainable through backpropagation for an end-to-end manner. The obtained features are reduced using PCA. In [22], authors introduce a spatial pyramid pooling (SPP) of CNN features which is an extension of the BoW. It generates a fixed-length representation regardless of image size/scale. Recently, authors in [59] introduce an end-to-end trainable network using multiscale local poling based on NetVLAD and a triplet mining. In [60], the authors present global descriptors REMAP based on a hierarchy of deep features using multiple CNN layers.

A number of methods have been developed to define binary codes based on deep learning [53,51]. Recently, a unified framework has been introduced for image retrieval and compression [51]. This framework applies the deep hashing method to learn compact binary codes and uses a new loss function to adapt

the binary representation. For retrieval purpose, VGG network is used with a specific configuration. For a manifold structure of the training data, K-Nearest Neighbor (KNN) algorithm is applied to create a neighborhood matrix during the learning of neural networks, which can be unsupervised or supervised. Experimental results show that this method outperforms some existing state-of-the-art ones.

Using CNN features can be global [41, 40], local [42, 44] or regional [43]. However, the CNN has a common architecture which described in the following section.

3 CNN architecture

Convolutional Layer: The convolutional (Conv) layer is the main component that allows extracting the images features. It is performed on the input image using a set of kernels (weights) as parameters. The different kernels are convolved across the width, height, and depth of the input volume using a dot product for returning the output volume. This layer comprises a rectangular grid of neurons. Specifically, each block of pixels is stretched into a matrix column, and the number of columns corresponds to the number of all local regions. As a result, the matrix multiplication is converted to the output volume with a depth that corresponds to the kernel number for obtaining the compact description of the input volume.

Pooling Layer: After each convolutional layer, a pooling layer may be used. The latter is a simple operation that is applied independently on the input volume. In the kernel, the pooling layer represents the outputs of neighboring groups of neurons. The most common type of pooling is the maximum. It is used to decrease the size (width and height dimensions) of the feature map while preserving the relevant information.

Normalization Layer: This layer supports a faster convergence. It allows adjusting the internal activations by using it before the activation function. In literature, various models of normalization have been proposed for ConvNet architectures. The two most commonly types used are the Local Response Normalization (LRN) and Batch Normalization (BatchNorm). The later performs a more global normalization. However, LRN implements the normalization in a small local neighborhood for each pixel. This method introduced in [40] and applied in [39] using the same parameters. The normalized output is given as following:

$$Y^i_{x,y} = X^i_{x,y} / \left(\kappa + \alpha \sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} (X^j_{x,y})^2 \right)^{\beta} \quad (1)$$

where $X^i_{x,y}$ and $Y^i_{x,y}$ are the pixel values at the (x, y) position of the the kernel i before and after normalization respectively. N is the total number of feature channels in X. The different constants are used as hyper-parameters where $\kappa = 2$, $n = 5$, $\alpha = 10^{-4}$ and $\beta = 0.75$.

Fully-Connected Layer: All neurons in this layer are connected to all activations in the previous layer, which can be calculated with a matrix multiplication followed by a bias offset. It takes as input the result of the previous layers (convolution and pooling) and returns a single vector which describe the image. However, this layer occupies the major part of CNN memory and requires high computation cost, due to the large number of parameters.

Correction layer (activation function): To improve the CNN efficiency, a correction layer is incorporated between layers. Its role consists in using an activation function to make the output nonlinear. The commonly used activation function is the rectified linear unit (ReLU) that applies an element wise function ($f(x) = \max(x, 0)$). In addition to its simplicity, the ReLu activation function does not require any additional parameter and does not change the size of input volume.

Loss layer: Loss layer is the last layer in the neural network. It specifies how network training penalizes the gap between expected and actual signal. Various loss functions, adapted to different tasks, can be employed. In particular, the Softmax function, also known as normalized exponential function, is used to predict a single class among K mutually classes. The Softmax function takes as input a vector of K real numbers and normalizes it into a probability distribution consisting of K probabilities ($\sigma(.)$) proportional to the exponentials of the input numbers. In the case of neural network, given the vector of the output layer $\mathbf{z} = [z_0, \ldots, z_{K-1}]$, the conventional Softmax function can be expressed as follows:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{2}$$

where j the index of the output unit, with $j = 1, 2, ..., K$.

Many studies [43,25,45] have shown that CNN features can be successfully retrieved from traditional LDR images. On the other hand, deep learning methodology has been previously proposed for companding HDR image from a single exposed LDR one [14,15,16]. Additionally, the CNN architecture has been used to reconstruct HDR video using multiple exposures captured over time [38]. But, to the best of our knowledge, the CNNs have never been investigated for the purpose of HDR image indexing and retrieval.

In this work, we attempt to exploit the many advantages of the CNNs to design an HDR image retrieval system. The proposed method is discussed in detail in the following section.

4 Proposed method

To determine the adequate descriptor, we model an HDR image as a collection of features using the VGG19 architecture [28]. Figure 1 summarizes the main steps that constitute the proposed scheme for Query-by-example HDR image retrieval. The database is divided into training and testing sets. Firstly, a perceptually uniform (PU) encoding [28] is applied on HDR images. This

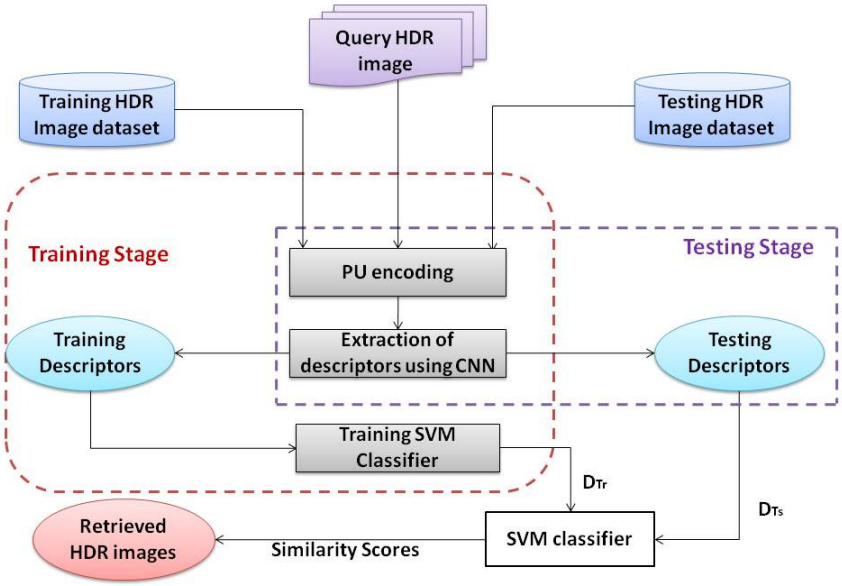

Fig. 1 Block diagram of the proposed HDR retrieval method.

encoding procedure is defined by a specific transfer function that prepares the HDR content to the feature extraction. Notably, the PU encoding is used to make sure that the distortion visibility is perceptually uniform through the coded pixels. For example, it can decrease the color sensitivity when the luminance is low. Secondly, a CNN descriptor is performed by extracting the rich features using FC layer activation. Doing so, each image in the dataset (training data and test data) will be indexed by a CNN feature vector (D_{Tr} (Descriptors of traing data) and D_{Ts}(Descriptors of testing data)). Finally, the sought HDR images are retrieved according to the Support Vector Machines (SVM) classification. We note that the model of SVM classifier is determined according to the features of training samples.

4.1 Extraction of CNN features

VGGNet [39] is a widely known network that encompasses 19 layers (convolutional and fully-connected). It is recognized by its simplicity and the large feature maps. Conforming to its architecture, each hidden layer uses the activation function ReLU. Depending on the training data, the size of RGB image is fixed to 224×224. A stack of convolutional layers is applied to the image using 3×3 filters, with very small sizes, for capturing the different left/right, up/down and center details like mentioned in [39]. For example, 1×1 convolution filters are used as linear transformation. The stride of convolution equals

to 1 pixel and the spatial padding is 1 pixel for 3×3 convolutional layers. Concerning the pooling layer, the operation is applied on 2×2 pixel window and stride 2. In total, five max-pooling layers are used after some convolutional ones. The convolutional layer is designated by Convm_n, where m and n refer to the order of the convolutional layer in the stack and the order of the stack, respectively. For example, Conv1_1 is the first convolutional layer in the first stack whereas Conv5_4 is the deepest layer in this network. In the top level of this architecture, there are three FC layers. The latter are applied after a set of convolutional layers that are characterized by the same architecture. The first and the second FC layers are of size 4096 channels; while the third one comprises only 1000 channels. As illustrated in Figure 2, the global network architecture uses 19 weight layers: 16 convolutional and 3 FC layers. It is worth noting that the VGG16 architecture uses a configuration with decreasing depth. Only 16 layers are implemented without using Conv3_4, Conv4_4 and Conv5_4 layers. In this network, each layer has different feature maps that can be applied as local descriptor with a specific dimension. Previous work established that the use of FC features improves the image retrieval accuracy[46,26]. This is mainly due to their high generalization and semantic descriptive ability. In this work, we propose to extract the 4096 dimensional output of the second FC layer. Then, replace the softmax layer with a linear SVM model which is recognized by its very good practical results. The next section gives an overview of the linear SVM.

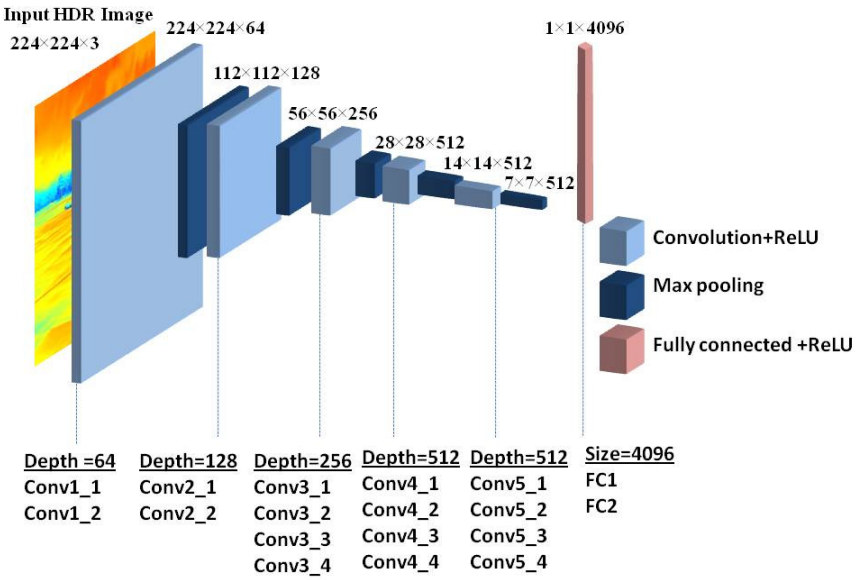

Fig. 2 The VGGNet architecture used for CNN features extraction

4.2 Linear SVM

SVM is one of the commonly used algorithms in machine learning applications, thanks to its powerful discriminative classifier. A hybrid approach which combines CNN and linear SVM has been proposed in some works such as [47, 48, 49]. The latters have concluded that CNN architecture can achieve an impressive performance if it is combined with linear SVM instead of Softmax. As demonstrated in [48], the linear SVM algorithm can be used for each layer with no additional fine-tuning of hidden representation.

In the field of image retrieval, the linear SVM algorithm allows to find the optimal separating hyperplane between classes using training samples. In our work, the feature vector is normalized used L_2. For a given training dataset x_i, y_i, with $i \in 1, \ldots, N$ and N the number of training samples, y_i equals $+1$ and -1 for class ω_1 and class ω_2, respectively. In the case of linearly separable descriptors, it is possible to find at least one hyperplane defined by a vector of weights w with a bias b, if we can separate the classes without error using the following classification function:

$$f(x) = w \cdot x + b = 0. \tag{3}$$

Therefore, to find an hyperplane, we need to estimate w and b using the following functions:

$$y_i(w \cdot x_i + b) \geq +1 \; for \; y_i = +1 \; (class \; \omega_1), \tag{4}$$

$$y_i(w \cdot x_i + b) \leq -1 \; for \; y_i = -1 \; (class \; \omega_2). \tag{5}$$

5 Experimental results

In this section, we present and compare the experimental results on LDR and HDR datasets. Firstly, we describe HDR databases and valuation criteria which were used in these experiments. Then, we provide a quantitative evaluation of the HDR image retrieval performance compared to LDR one using the features extracted with the CNN framework. Also, we provide comparative evaluations to other related methods. Finally, we assess the time complexity.

5.1 Databases and measures

Although there is a growing interest in HDR content, the amount of available data remains limited to evaluate and test HDR indexing and retrieval systems. Fortunately, the rapid development of HDR tools has made it possible to easily generate an HDR image. Thus, in order to create HDR databases, we used the inverse tone mapping method presented in [13], which provides very satisfactory results, to build HDR images by expending LDR ones. In the current work, we consider LDR PASCAL VOC2007, CIFAR-10, and Wang databases.

– PASCAL VOC2007 database: It is one of the most widely used benchmark for image classification. It contains 9963 RGB images divided into 20 classes (Person, Bird, Cat, Cow, Dog, Horse, Sheep, Airplane, Bicycle, Boat, Bus, Car, Motorbike, Train, Bottle, Chair, Dining table, Potted plant, Sofa and TV/Monitor) and is available at http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html

– CIFAR-10: It is one of the most popular deep learning dataset. It comprises 60000 natural images of size 32×32 divided into 10 categories (Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, and Truck). From the CIFAR-10 collection, 50000 images are devoted to training (5000 for each class) while the remaining 10000 images are devoted to testing (1000 for each class) https://www.cs.toronto.edu/~kriz/cifar.html

– Wang: It contains 1000 images classified into ten categories (Africa, Beach, Buses, Monuments, Dinosaurs, Elephants, Flowers, Horses, Mountains, Food). Each category comprises 100 images. http://wang.ist.psu.edu/docs/related/

Several performance measures can be used to asset the efficiency of indexing/retrieval methods. Specifically, evaluation over different datasets is performed by using testing images (one or more images as the query), and ranking the images from the most similar to the least similar. The performance for a particular method is estimated by the average of the performances overall query images. To asset the efficiency of the test methods we retained the following measures:

– Precision against recall plot: A curve illustrating the relationship between precision and recall for retrieval system. The precision represents the ability of retrieval algorithm to return only images that are relevant whereas the recall corresponds to the system ability to return all images that are relevant.
– mAP: The mean Average Precision (mAP) of a set of queries is a common metric used to evaluate the effectiveness of an image retrieval system. It is worth noting that among retrieval evaluation measures, mAP has been shown to have good discrimination and stability.
– Accuracy: It can be defined as the percentage of correctly classified instances

5.2 Image retrieval Results

To evaluate the performance of our HDR image retrieval approach and investigate the impact of CNN, retrieval experiments are carried out from HDR and LDR versions of PASCAL VOC2007 database using the mAP scores. Table 1 shows the retrieval accuracy using SIFT and CNN descriptors for LDR (LDR-SIFT/LDR-CNN), Expanded-Mapped (EM-SIFT), HDR with lin-

ear luminance values (HDR-Lin-CNN) and HDR with PU encoding (HDR-PU-SIFT/HDR-PU-CNN) representations. We note that both LDR-SIFT and EM-SIFT [35] methods use a bag of visual words as descriptor.

From the results reported in Table 1, we observe that in the case of HDR-PU-CNN descriptor, the mAP scores reveal the excellent results for the majority of classes. This may be explained by the fact that PU encoding plainly improves the effectiveness of the CNN descriptor. Additionally, this encoding procedure leads to more accurate results than that provided by original LDR content. On average, PU encoding provides a gain of about 2.71% and 1.11% for CNN and SIFT features respectively, when compared to the same features obtained from LDR representation.

Table 1 mAP scores on the HDR PASCAL VOC2007 dataset using VGG19

Classes	**Descriptors**					
	LDR-SIFT	***EM-SIFT***	***HDR-PU-SIFT***	***LDR-CNN***	***HDR-Lin-CNN***	***HDR-PU-CNN***
Airplane	87.83	88.65	88.48	99.08	98.49	99.13
Bicycle	64.12	67.70	68.13	96.15	92.90	96.29
Bird	73.27	74.38	73.67	98.30	96.31	98.98
Boat	77.45	79.72	80.98	98.80	95.78	98.28
Bottle	39.66	38.83	37.96	70.93	61.67	74.25
Car	90.25	90.31	91.90	99.39	98.02	99.17
Cat	71.07	71.28	72.78	90.81	91.13	94.88
Dog	68.57	68.44	71.20	93.60	87.81	93.90
Horse	70.71	73.01	71.86	97.46	87.77	97.48
Motorbike	74.84	75.25	76.18	98.12	96.20	98.63
Person	85.45	85.64	85.70	95.37	93.20	95.07
Pottedplant	33.17	34.53	34.31	63.95	58.06	62.65
Sheep	45.27	46.10	46.28	94.46	85.06	92.60
Train	81.44	88.40	78.93	99.14	96.74	99.36
TVmonitor	48.80	48.92	49.04	75.65	65.68	77.98
Average	66.21	66.96	67.32	87.96	85.79	90.67

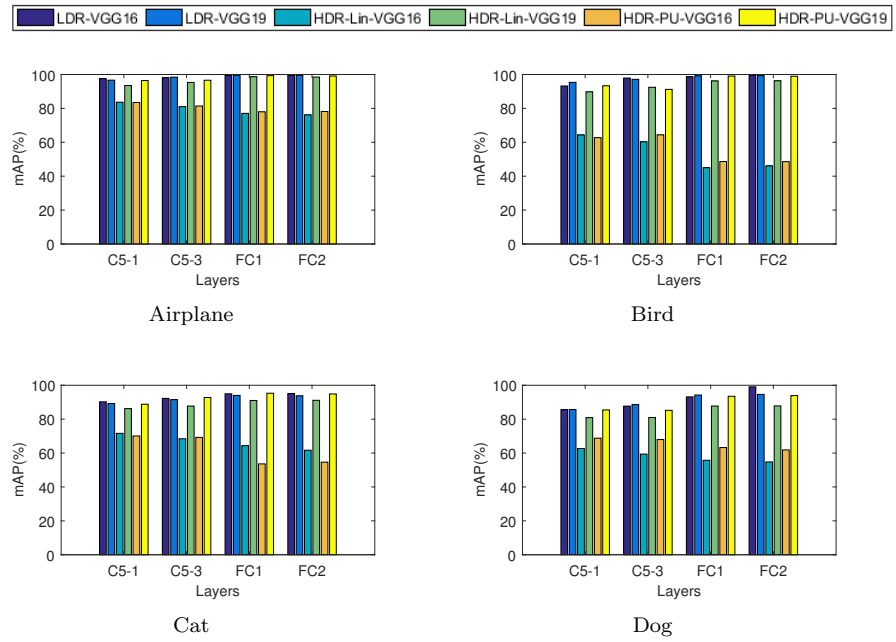

Fig. 3 Retrieval performance using mAP measure for LDR, HDR-Lin and HDR-PU representations using VGG16 and VGG19 frameworks on some classes of PASCAL VOC2007 dataset.

Earlier studies have shown that the SIFT descriptor is recognized by its ability to capture the local object details like edges and corners and, consequently, achieves a good performance for image retrieval. In the case of HDR images, our experiments prove that the use of PU encoding allows to enhance the representation of HDR images. On the basis of the results reported in Table 1, the majority of mAP values of HDR-PU-SIFT are higher than 70%. However, in some classes like Bottle, Pottedplant and Sheep, the mAP scores are lower than 50%. In counterpart, by examining the results obtained with CNN descriptor, it appears that the later entails advantages in terms of retrieval efficiency thanks to its capability to learn different images and successfully model HDR data. The results of our analysis reported in Table 1 clearly show that the CNN descriptor systematically outperforms EM-SIFT [35] and HDR-PU-SIFT ones. Moreover, we remark that the mAP scores obtained by HDR-PU-CNN method surpass 95% for most classes.

From a more quantitative point of view, we studied the impact of the extracted features using VGG16 and VGG19 on LDR, HDR-Lin (HDR with linear luminance values) and HDR-PU (HDR with PU encoding) representations. For a given HDR-PU content, we extracted features from some layers (conv5_1 (C5_1), conv5_3 (C5_3), FC1 and FC2) and compared the matching precisions to those obtained for LDR and HDR-PU versions in PASCAL

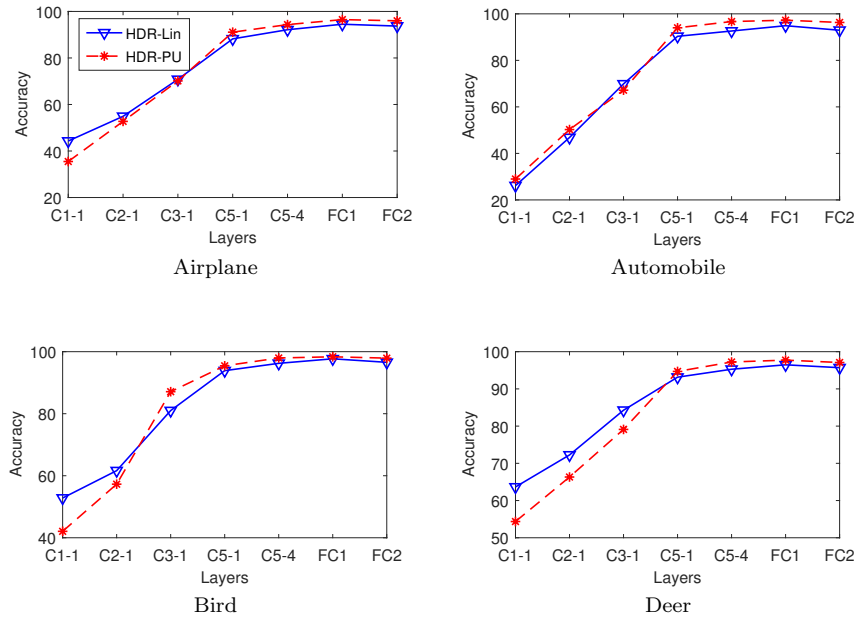

Fig. 4 Accuracy results for some classes of CIFAR-10 database, using different layers in VGG19 framework.

VOC2007 dataset. From Figure 3, we observe that the VGG19 model achieves higher performance for HDR content (HDR-Lin and HDR-PU) in different layers. It gives very high precision and outperforms the VGG16 model. We believe that this is mainly due to the number of layers which strongly influences the richness of the information in the HDR feature descriptor. However, in the case of LDR content, the results obtained by using VGG16 and VGG19 are almost similar. Again, one may also conclude that PU encoding improves the precision of the HDR content and consequently induces an overall retrieval efficiency gain. For instance, the gain of PU is about 20.92% in Airplane class using FC2 as descriptor. Moreover, we can clearly notice that when FC layers are considered, the mAP scores for LDR and HDR-PU contents are very close. For some classes like Dog, the LDR content exhibits a superior performance compared to HDR representation. This limitation stems from the high sensitivity of HDR content that badly affects the matching accuracy against the original dataset. As discussed above, the PU encoding has proven its usefulness to alleviate this limitation by adjusting the HDR pixel values and reduce their sensitivity.

Table 2 Comparison Accuracy scores with other methods using LDR and HDR represenations on the Wang dataset

Global Method [31]		**Local Method [35]**			**CNN Method**	
LDR	***HDR***	***LDR***	***HDR***	***Expended-mapped***	***HDR***	***LDR***
68.49	68.62	92.34	57.78	93.16	100	100

Table 3 Performance comparison (mAP) with other method on the Cifar-10 dataset

Method	***mAP***
Proposed method (LDR)	94.6
Proposed method (HDR-PU)	95.1
BGAN+ [51]	89.4

We have also tested our retrieval method, using VGG19 network, on the CIFAR-10 database. Figure 4 presents the accuracy results for some classes. From this figure, we can see that the FC features outperform the convolutional ones for all the recall values. Specifically, the accuracy of the top layers is superior to that obtained from the bottom ones. One may also notice that in the case of C5-1 layer, the PU encoding shows a distinctive improvement for all the classes. Quantifying the retrieval performance improvements, brought by PU encoding, a gain in accuracy of about 3.35% and 1.36% is attained for Automobile and Deer classes, respectively, when compared it to the HDR-Lin using descriptor FC1.

In Table 2, we compare the CNN method for HDR and LDR representations with other state-of-the-art methods (Global and local) on Wang dataset. 30 images per class are randomly selected for training. From this table, we can notice that the CNN descriptor offers the best result for both LDR and HDR content. It achieves good retrieval power for HDR content because the features are more descriptive than those extracted from [35] and [31]. This is explained by the fact that CNN descriptor fits well for HDR content. Meanwhile, local method(SIFT) owns the worst accuracy.

Table 3 shows the CIFAR-10 retrieval results based on the mAP for the proposed method and BGAN+ [51], a recent state-of-the-art learning-based hashing method for image retrieval. According to these results, we observe that the proposed method achieve strong results, and it outperforms BGAN+ in terms of precision.

As a last comparison, we carried out evaluations of the proposed retrieval method on the whole CIFAR-10 database. Figure 5 illustrates the precision–recall plots for the whole dataset. From the depicted curves, we observe that the configuration that uses the first and second FC layers owns a higher performance against the other ones. However, the performance decreases when using the bottom layers of the CNN framework. One may safely conclude that the selection of the network architecture and depth are crucial to achieve a successful HDR content retrieval.

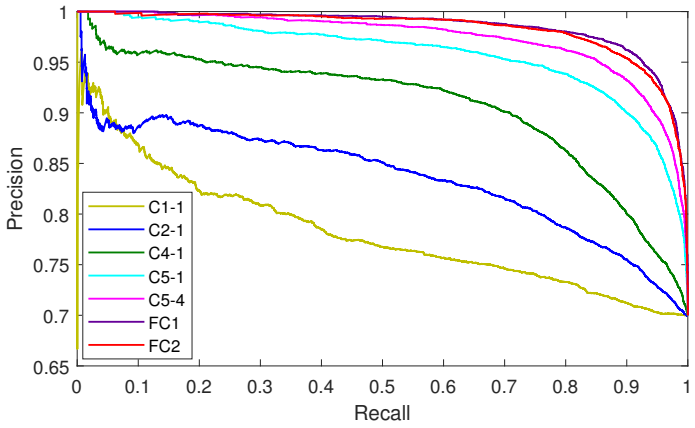

Fig. 5 Precision vs Recall curves of the tested VGG19 framework (under different configurations) using the HDR-PU representation on the whole of CIFAR-10 database.

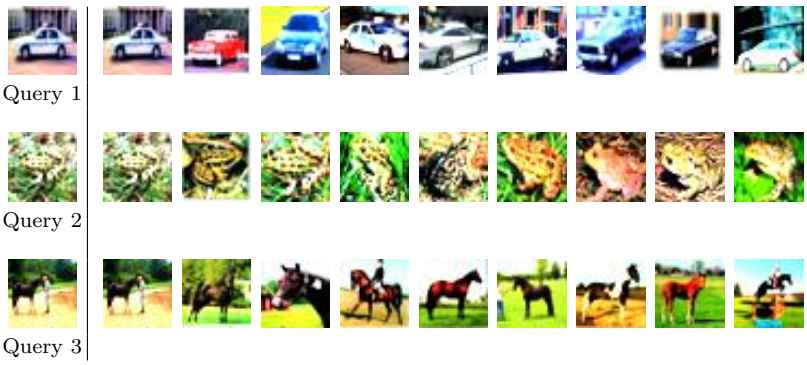

Fig. 6 Some examples of query HDR images from our HDR CIFAR-10 dataset and the top-9 retrieval results

Figure 6 shows three queries and their corresponding top 9 retrieved HDR images from CIFAR-10 dataset using the proposed method. Query 1 is from Automobile class, Query 2 is from Frog class and Query 3 is from Horse class. The different HDR images are tone mapped using the TMO proposed in [31] to ensure the rendering on LDR devices. As we can see from this figure, the retrieved images in the first positions of the rank lists belong to the same class as their corresponding queries. This proves the effectiveness of the CNN features in HDR image retrieval despite the large scale difference between images.

5.3 Complexity evaluation

In order to evaluate the time complexity of the proposed CNN method, execution-time tests are performed on a machine with an Intel(R) Xeon CPU E5-2640

v4 Core 10 at 2.40 GHz. For example, for the CIFAR-10 dataset, the running time for the retrieval of an HDR image takes on average: 400 ms for CNN Features Extraction, 4 ms for PU encoding, and 78 ms for SVM classifier.

6 Conclusion

We have presented an HDR image retrieval method based on the CNN paradigm. To ameliorate the accuracy of the proposed approach, the PU encoding is applied on the HDR pixel values before passing to the computation of the descriptor components. Through this work, we have reported, for the first time, results competing with some methods on the challenging HDR PASCAL VOC2007, CIFAR-10 and Wang datasets. In the same context, we have provided good practices for extracting features from some layer of the CNN, using VGG19 pre-trained model. Experimental assessments have demonstrated that the CNN features exhibit substantial performance improvements over SIFT descriptor. Moreover, the obtained results reveal that the FC layers offer the best performance among the other CNN intermediate layers. Hence, we can claim that it is very appropriate for describing HDR images. However, it must be emphasized that in some layers, especially in VGG16 framework, the accuracy of our retrieval method, applied on HDR images is lower than its counterpart applied on LDR images. This discomfort is particularly faced when using a decreasing number of layers. As future work, we intend to resolve this problem, by incorporating additional discriminative cues in the layers to enhance the HDR features based on CNN descriptor. As another promising line of future work, we plan to investigate the HDR datasets on deep CNN architecture like ResNet and adjust it to HDR content.

References

1. Dufaux, F., Callet, P. L., Mantiuk, R., Mrak, M.: High Dynamic Range Video: From Acquisition, to Display and Applications. Academic Press (2016)
2. Chalmers, A., Debattista,K.: HDR Video Past, Present and Future: A Perspective. Signal Processing: Image Communication, 54, 49–55, (2017)
3. Mantiuk, R. K., Myszkowski, K. H., Seidel, P.: High Dynamic Range Imaging. Wiley Encyclopedia of Electrical and Electronics Engineering, pp. 1–4, (2015)
4. Debevec, P. E., Malik, J.: Recovering high dynamicrange radiance maps from photographs. Proceeding of SIGGRAPH, pp. 369–378, (1997)
5. Mitsunaga, T., Nayar, S. K.: Radiometric self calibration. Conference on Computer Vision and Pattern Recognition, pp. 374–380, (1999)
6. Ward, G.: Real pixels. Graphics Gems, New York (1991)
7. Larson, G. W.: Logluv encoding for full-gamut, high-dynamic range images. Journal of Graphics Tools, 3(1), 15–31 (1998)
8. OpenEXR, http://www.openexr.org (2003)
9. Kim, BK., Park, RH., Chang, S.: Tone mapping with contrast preservation and lightness correction in high dynamic range imaging. Signal, Image and Video Processing, 10(8), 1425–1432, (2016)
10. Banterle, F., Ledda, P., Debattista, K., Chalmers, A.: Inverse tone mapping. international conference on Computer graphics and interactive techniques, pp. 349–356, (2006)

11. Masia, B., Serrano, A., Gutierrez, D.: Dynamic range expansion based on image statistics. Multimedia Tools and Applications, 76(1), 631–648 (2017)
12. Kovaleski, R. P., Oliveira, M. M.: High-quality brightness enhancement functions for real-time reverse tone mapping. The Visual Computer, 25(5), 539–547 (2009)
13. Kovaleski, R. P., Oliveira, M. M.:High-Quality Reverse Tone Mapping for a Wide Range of Exposures. Conference on Graphics, Patterns and Images, pp. 49–56, (2014)
14. Zhang, J., Lalonde, J. F.: Learning High Dynamic Range from Outdoor Panoramas. International Conference on Computer Vision, (2017)
15. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R. K., Unger, J.: HDR image reconstruction from a single exposure using deep CNNs. ACM Transactions on Graphics, 36(6), 178:1–178:15 (2017)
16. Endo, Y., Kanamori, Y., Mitani, J.: Deep Reverse Tone Mapping. ACM Transactions on Graphics, 36(6) (2017)
17. Zhang, S., He, F., Ren, W.: NLDN: Non-local Dehazing Network for Dense Haze Removal. Neurocomputing, 410, 363–373 (2020)
18. Zhang, S., He, F.: RCDN: Learning Deep Residual Convolutional Dehazing Networks. The Visual Computer, 36(9), 1797–1808 (2020)
19. Pan, Y., He, F., Yu, H.: Learning social representations with deep autoencoder for recommender system. World Wide Web, 23, 2259–2279 (2020)
20. Quan, Q., He, F., Li, H.: A multi-phase blending method with incremental intensity for training detection networks. The Visual Computer,(2020)
21. Gao, L., Li, X., Song, Shen HT, J,: Hierarchical LSTMs with Adaptive Attention for Visual Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(5), 1112–1131 (2020)
22. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. European Conference on Computer Vision, pp. 346–361, (2014)
23. Gong Y., Wang L., Guo R., Lazebnik S.: Multi-scale orderless pooling of deep convolutional activation features. European Conference on Computer Vision, pp. 392–407, (2014)
24. Razavian, A. S., Sullivan, J., Carlsson, S., Maki, A.: Visual instance retrieval with deep convolutional networks. ITE Transactions on Media Technology and Applications, 4(3), 251–258 (2016)
25. Babenko, A., Lempitsky, V.: Aggregating deep convolutional features for image retrieval, International Conference on Computer Vision, pp. 1269–1277 (2015)
26. Kalantidis, Y., Mellina, C., Osindero, S.: Cross-dimensional weighting for aggregated deep convolutional features. European Conference on Computer Vision, pp. 685–701, (2016)
27. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. International conference on learning representations, pp. 1–12, (2016)
28. Aydin, T. O., Mantiuk, R., Seidel, H. P.: Extending Quality Metrics to Full Luminance Range Images. Human Vision and Electronic Imaging XIII (Proceedings of SPIE), (2008)
29. Khwildi, R., Hachani, M., Ouled Zaid, A.: New indexing method of HDR images using color histograms. International conference on machine vision, (2016)
30. Khwildi, R., Ouled Zaid, A.: Color Based HDR Image Retrieval Using HSV Histogram and Color Moments. International Conference on Computer Systems and Applications, pp. 1–5, (2018)
31. Khwildi, R., Ouled Zaid, A.: HDR image retrieval by using color-based descriptor and tone mapping operator. The Visual Computer, 36, 1111–1126 (2020)
32. Rana, A., Valenzise, G., Dufaux, F.: An Evaluation of HDR Image Matching under Extreme Illumination Changes. Visual Communications and Image Processing, pp. 1–4, (2016)
33. Rana, A., Valenzise, G., Dufaux, F.: Evaluation of feature detection in HDR based imaging under changes in illumination conditions. IEEE International Symposium on Multimedia, pp. 289–294, (2015)
34. Bronislav, P., Chalmers, A., Zemcík, P., Hooberman, L., Zadík, M.: Evaluation of feature point detection in high dynamic range imagery. Journal of Visual Communication and Image Representation, 28(C), pp. 141–160 (2016)

35. Khwildi, R., Ouled Zaid, A.: New retrieval system based on low dynamic range expansion and SIFT descriptor. International Workshop on Multimedia Signal Processing, pp. 1–6, (2018)
36. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural Codes for Image Retrieval. European Conference on Computer Vision, pp. 58–599, (2014)
37. Razavian, A. S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. Computer Vision and Pattern Recognition Workshops, pp. 512–519, (2014)
38. Kalantari, N. K., Ramamoorthi, R.: Deep High Dynamic Range Imaging of Dynamic Scenes. ACM Transactions on Graphics, 36(4), 144:1–144:12 (2017)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, (2015)
40. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90 (2017)
41. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. European Conference on Computer Vision, pp. 584–599, (2014)
42. Ng, J., Yang, F., Davis, L.: Exploiting local features from deep networks for image retrieval. Conference on Computer Vision and Pattern Recognition Workshops, pp. 53–61, (2015)
43. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. European Conference on Computer Vision, pp. 392–407, (2014)
44. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. European European Conference on Computer Vision, pp. 834–849, (2014)
45. Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: Visual instance retrieval with deep convolutional networks. International Conference on Learning Representations, (2015)
46. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral maxpooling of CNN activations. International Conference on Learning Representations, (2016)
47. Tang, Y.: Deep learning using linear support vector machines. International Conference on Neural Information Processing, pp 458–465, (2013)
48. Vinyals, O., Jia, Y., Deng, L., Darrell, T.: Learning with Recursive Perceptual Representations. Annual Conference on Neural Information Processing Systems, pp. 2834–2842, (2012)
49. Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Jiao, J.: Orientation robust object detection in aerial images using deep convolutional neural network. International Conference on Image Processing, pp. 3735–3739, (2015)
50. Lee, C. Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-Supervised Nets. International Conference on Artificial Intelligence and Statistics, (2015)
51. Song, J., He, T., Gao, L. et al: Unified Binary Generative Adversarial Network for Image Retrieval and Compression. International Journal of Computer Vision volume, 128, 2243–2264 (2020)
52. Song, J., Zhang, H., Li, X., et al: Self-Supervised Video Hashing With Hierarchical Binary Auto-Encoder. IEEE Transactions on Image Processing, 27(7), 3210–3221 (2018)
53. Lin, K., Lu, J., Chen, C., Zhou, J.: Learning Compact Binary Descriptors with Unsupervised Deep Neural Networks. Conference on Computer Vision and Pattern Recognition, pp. 1183–1192, (2016)
54. Zheng, L., Zhao, Y., Wang, S., Wang, J., Tian, Q.: Good practice in cnn feature transfer. arXiv preprint arXiv:1604.00133, (2016)
55. Mohedano, E., McGuinness, K., et al: Bags of local convolution. International Conference on Multimedia Retrieval, pp. 327–331, (2016)
56. Uricchio, T., Bertini, M., Seidenari, L., Del Bimbo, A.: Fisher encoded convolutional Bag-of-Windows for efficient image retrieval and social image tagging. International Conference on Computer Vision Workshop, pp. 1020–1026, (2015)
57. Radenovic, F., Tolias, G., Chum., O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7), 1655–1668 (2018)

58. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision, 124(2), 237–254 (2017)
59. Vaccaro, F., Bertini, M., Uricchio, T., Del BimboImage, A.: Retrieval using Multi-scale CNN Features Pooling, International Conference on Multimedia Retrieval, pp. 311–315, (2020)
60. Husain, S. S., Bober, M.: Multi-Layer Entropy-Guided Pooling of Dense CNN Features for Image Retrieval. IEEE Transactions on Image Processing, 28(10), 5201–5213 (2019)