Signal Processing: Image Communication 80 (2020) 115652

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Robust image retrieval by cascading a deep quality assessment network ☆

Biju Venkadath Somasundaran [a], Rajiv Soundararajan [a,*], Soma Biswas [b]

[a] Department of ECE, Indian Institute of Science, Bangalore, 560012, India
[b] Department of EE, Indian Institute of Science, Bangalore, 560012, India

ARTICLE INFO

Keywords:
Image enhancement
Image quality assessment
Deep convolutional neural network
Denoising
Super resolution
Image retrieval

ABSTRACT

The performance of computer vision algorithms can severely degrade in the presence of a variety of distortions. While image enhancement algorithms have evolved to optimize image quality as measured according to human visual perception, their relevance in maximizing the success of computer vision algorithms operating on the enhanced image has been much less investigated. We consider the problem of image enhancement to combat Gaussian noise and low resolution with respect to the specific application of image retrieval from a dataset. We define the notion of image quality as determined by the success of image retrieval and design a deep convolutional neural network (CNN) to predict this quality. This network is then cascaded with a deep CNN designed for image denoising or super resolution, allowing for optimization of the enhancement CNN to maximize retrieval performance. This framework allows us to couple enhancement to the retrieval problem. We also consider the problem of adapting image features for robust retrieval performance in the presence of distortions. We show through experiments on distorted images of the Oxford and Paris buildings datasets that our algorithms yield improved mean average precision when compared to using enhancement methods that are oblivious to the task of image retrieval. [1]

1. Introduction

The proliferation of smart mobile devices has led to an explosion in the amount of images that are captured, stored and analyzed. On the other hand, the availability of increased compute power and internet connectivity has enabled the application of sophisticated computer vision algorithms for visual analytics. Indeed, the fruits of such advances have resulted in applications such as Google Lens which can improve the quality of lives of humans by providing a wealth of information. However, the performance of computer vision algorithms on camera captured images can degrade due to a variety of distortions such as noise, resolution, compression and illumination. In order to provide a reliable extraction of visual analytics, there is a need to ensure robustness of the computer vision algorithms in the presence of such distortions. In this paper, we focus on a specific instance of this robustness question by considering the problem of image retrieval. We consider the design of image enhancement algorithms to ensure the robust performance of retrieval algorithms in the presence of distortions due to noise and low resolution. We note that the image retrieval algorithm we refer to here is the classical retrieval problem where the goal is to retrieve images from a database with similar content or semantic similarity.

Image retrieval based on the bag of words model has been studied quite extensively [1,2]. Several improvements have also been proposed to overcome the limitations of feature detectors and descriptors, descriptor comparison metrics and quantization of descriptors [3–5]. Nevertheless, the performance of image retrieval in the presence of distortions and how to improve performance in such scenarios has been much less studied. Image denoising and super resolution are problems with rich literature and successful algorithms have been developed. Various techniques developed over the years have evolved to optimize the perceptual quality of the enhanced images. Improved statistical priors on natural images and the idea of exploiting the similarity of patch content across the image have led to image denoising algorithms with excellent performance [6,7]. The theory of sparse signal representations has been used to develop state of the art single image super resolution algorithms [8]. Recently, deep convolutional neural networks (CNN) have been successfully deployed for both image denoising and super resolution [9]. It is shown that state of the art performance can be achieved for both these problems using simple architectures of CNNs. While all these algorithms lead to images with very good perceptual

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.image.2019.115652.

* Corresponding author.
E-mail addresses: bijuvselankur@gmail.com (B.V. Somasundaran), rajivs@iisc.ac.in (R. Soundararajan), somabiswas@iisc.ac.in (S. Biswas).

[1] The material in this paper appeared in part at the 2018 IEEE International Conference on Image Processing, Athens, Greece.

https://doi.org/10.1016/j.image.2019.115652
Received 7 February 2019; Received in revised form 28 September 2019; Accepted 28 September 2019
Available online 5 October 2019

quality, their relevance to the success to computer vision algorithms and in particular, image retrieval, has been much less studied.

At first sight, the optimization of classical denoising and/or single image super resolution algorithms for image retrieval tasks appears to be challenging. This is partly because the denoising or super resolution algorithms themselves are complex involving non-linear operations of various parameters that need to be optimized. While the use of deep CNNs simplifies the enhancement operation to some extent, networks are typically optimized for cost functions such as regularized mean squared error or perceptual quality indices such as the structural similarity index [10]. While these cost functions may be relevant for perceptual quality, their relevance in improving the performance of image retrieval is not clear. The measurement of image retrieval performance involves two components, the retrieval algorithm itself and the performance evaluation of the output of the retrieval algorithm in terms of metrics such as average precision by comparing the output with an annotated database. These involve a complex sequence of operations that cannot be written as a closed form expression. Thus it is not clear how a differentiable cost function can be obtained that can be used to optimize the image enhancement algorithms.

Our main contribution is in the design of a framework for image denoising and super resolution for image retrieval. We first design a deep CNN to predict the image retrieval performance in terms of average precision as a function of image distortions. We refer to this CNN as the quality assessment for image retrieval (QAIR) CNN since it predicts the image quality as relevant to image retrieval. We then cascade this CNN to the output of a denoising or super resolution CNN and use the output of the QAIR CNN to optimize the weights of the denoising or super resolution CNN through back propagation. This architecture provides a seamless method to optimize the denoising or super resolution for improving image retrieval performance. We conduct experiments to show that the QAIR CNN is efficient in predicting the image quality of the distorted image. Further we also show that by coupling the enhancement CNN with the QAIR CNN, we are able to improve the performance of image retrieval when compared to approaches which treat enhancement and retrieval as separate problems.

In contrast to the approach of image enhancement to achieve robust image retrieval, we then consider the complementary problem of feature adaptation for image retrieval. The goal of this problem is to design a framework that allows the learning of features for the image retrieval task at hand in the presence of distortions. Further, while we seek to learn features, the rest of the pipeline in the given retrieval algorithm remains unchanged. The features will need to be learnt appropriately for a different retrieval task. While a generic solution appears to be challenging, we present a solution for adapting deep CNN based features used for image retrieval [11]. In particular, we design a QAIR CNN which takes as input, the deep CNN based features in [12] and predicts the average precision of the image retrieval. We then cascade this QAIR CNN with the deep CNN used to generate the features to fine tune the later CNN while keeping the former fixed. We show that this feature adaptation leads to an improvement in performance of the deep CNN based features with respect to noise and low resolution.

We published preliminary results of our work in a conference version which only focussed on the problem of image denoising for image retrieval for specific noise levels [13]. In this paper, we also consider the complementary problem of feature adaptation for robust image retrieval on distorted images and show how our framework can be used to solve this problem as well. This material is contained in Section 5 and is completely new. We also extend our image denoising framework for a set of noise levels instead of individual noise levels. Further, we apply our framework to perform image super resolution for image retrieval. The extension to super resolution is discussed in Section 4.2. The experimental results corresponding to all the new material are contained in Sections 6.3, 6.4.3, 6.6, 6.7, and 6.9.

The rest of the paper is organized as follows. In Section 2, we present an overview of the related work. We describe our method of quality assessment for image retrieval in Section 3, the image enhancement framework in Section 4 and the feature enhancement approach in Section 5. We present detailed experiments and comparisons in Section 6 and conclude the paper in Section 7.

2. Related work

We now discuss prior work related to our problem. We identify five different areas in image retrieval, image denoising, super resolution, quality assessment and the connection between computer vision algorithms and image quality as related to our work. We discuss these in the following.

Image retrieval usually refers to the problem of retrieving a set of images relevant to a query image containing a particular object. Successful retrieval algorithms based on the construction of a bag of visual words have been developed [1,2]. Several researchers have improved the performance of retrieval algorithms by designing different feature descriptors [4]. Further, spatial and geometric constraints [14,15] have also led to improved performance. Compact codes have been designed based on local image descriptions to speed up the retrieval algorithms [16]. While majority of the approaches deal with improvements in the image retrieval pipeline, the robustness of the retrieval algorithm to image quality degradations such as noise and resolution has been much less studied.

There is rich literature in image denoising. One of the state of the art denoising methods is Block-Matching and 3D Filtering (BM3D) [6], which is based on non-local self similarity and combines multiple steps such as block matching, collaborative filtering on different blocks and aggregation of different blocks to form the denoised image. Other successful image denoising algorithms such as those based on expected patch log likelihood (EPLL) [7] and Gaussian scale mixture models [17], explore the availability of rich natural scene statistical models. Sparse representations of images have also led to successful image denoising algorithms [18]. While neural networks were initially explored for image denoising [19], deep convolutional neural networks (CNNs) such as DnCNN [9] and FFDNet [20] have been shown to achieve state of the art image denoising performance.

The problem of image super resolution has also been addressed by several researchers. Improving resolution by image registration [21] and example based super resolution [22] are examples of super resolution using multiple low resolution images. One of the earlier pieces of work on single image super resolution was done by Glasner et al. [23] by exploiting the recurrence of patches in an image, both at the same scale as well as across scales. Dong et al. designed a CNN called SRCNN which had 3 layers and achieves super resolution on image patches [24]. Kim et al. came up with a deep CNN based model for image super resolution [25] inspired by the VGG-net for image classification by predicting the residual image given an up sampled low resolution image. This residual image is then added to the up sampled image to generate the high resolution image. DnCNN [9] also adopts a similar approach to solve the super resolution problem.

The problem of perceptual image quality assessment has rich literature and significant progress has been made on no reference image quality assessment through algorithms such as DIVIINE [26], BLIINDS [27], BRISQUE [28] and CORNIA [29]. While the above algorithms operate based on natural scene statistics based features, there have been several efforts based on convolutional neural networks [30–33]. In [31], a pre-trained deep CNN to extract image features is combined with dense fully connected layers to predict perceptual image quality. CNN based architectures have been also been applied successfully in both full reference and no reference QA through a unified framework [34].

The impact of image quality on computer vision tasks has been much less studied. Perceptual image quality features are shown to be relevant for robust face detection [35]. The notion of machine vision quality is used to design image enhancement algorithms for

face detection [36]. The relation between image quality and image utility or the usefulness of an image with respect to performing a particular task is explored in [37]. The relation between image quality and the performance of object tracking has also been studied [38]. More recently, image denoising algorithms have been optimized for a deep learning based image classification problem [39].

3. Image quality assessment for image retrieval

In this section, we define the notion of image quality with respect to the success of the specific computer vision task of image retrieval. We first describe the performance measurement of image retrieval and then define our notion of quality for image retrieval. An image retrieval algorithm takes as input, an image database and a query image and returns as output, matching images from the database in order of their similarity to the query image. An example of a retrieval algorithm based on the scale invariant feature transform (SIFT) is shown in Fig. 1. Image retrieval involves the computing of image features and their comparison with a database of images subject to some geometric consistency checks. Thus, the output of the retrieval algorithm is a complex function of the input image. While we present an example based on the SIFT features above, our framework applies to any image retrieval algorithm in general.

3.1. Image quality index

We define the quality of an image for image retrieval in terms of the success of the retrieval task. In particular, we define quality as the average precision achieved on a given test image with respect to the database [1]. Mathematically, let precision and recall of retrieval be defined as

$$Precision = \frac{CM}{RI}, Recall = \frac{CM}{TM}, \tag{1}$$

where RI is the number of retrieved images for a query image, CM is the number of correct matches in the set of retrieved images and TM is the total number of true matches in the database for that query image. The number of images in the sorted list output by the retrieval algorithm can be varied using a threshold to obtain a precision–recall curve. We define image quality as the average precision or the area under the precision–recall curve. Note that the average precision that we seek to predict is a function of the given retrieval algorithm.

Before we present algorithms for predicting image quality for image retrieval, we discuss how this notion of image quality can be different from perceptual image quality, which is typically associated with a task free viewing condition and human perception. The example in Fig. 2 shows the difference between quality assessment for image retrieval and perceptual quality assessment. An image which looks visually good may not give good results when used for image retrieval. On the other hand, an image which has visible distortions may yet be good from the point of view of the success of image retrieval. Thus the relation between the presence of distortions in an image and the success of a computer vision task is complex and needs to be learnt carefully.

3.2. Image QA CNN

Having defined the notion of image quality with respect to image retrieval, we now consider the problem of designing algorithms to predict this quality given a potentially distorted image. We design a CNN to predict this quality directly from the image. The use of a CNN instead of specific features such as those in [26,28] for image retrieval QA is motivated by their suitability for optimizing image enhancement as discussed in Section 4. Since we do not have enough data to train a CNN for this purpose from scratch, we use pre-trained convolutional layers of the VGG-16 CNN [12] trained for image classification on the ImageNet dataset [40], and augment it with 5 fully connected layers at the end. This is similar in nature to the approach in [31] to predict perceptual image quality. The pre-trained CNN is shown in Fig. 3 and the fully connected layers are shown in Fig. 4. The first fully connected layer has 128 nodes and the last layer has a single node corresponding to the output. All the layers except the last layer have rectified linear units (ReLU) as activation functions. Initially, the convolutional layers are frozen and the only the fully connected layers are trained using Adam optimizer. After sufficient training, the last 9 convolutional layers of the VGG-16 network are unfrozen and fine tuned using stochastic gradient descent (SGD) optimizer with a low learning rate of 10^{-3}. We refer to our CNN architecture as the quality assessment for image retrieval (QAIR) CNN.

We divide the image into patches of size 124×124 and train the QAIR CNN on image patches to predict the average precision of the distorted image from which these patches are drawn. Let x_n and y_n be the ground truth and predicted quality scores (or average precision) of the nth image patch and let N be the total number of patches. Then for training the QA CNN, we use the mean absolute error loss function defined as,

$$L = \frac{1}{N}\sum_{n=1}^{N} |y_n - x_n|. \tag{2}$$

4. Image enhancement framework for image retrieval

We now describe our approach to image enhancement for image retrieval. We consider two different image enhancement scenarios for image retrieval, image denoising and image super resolution. Our goal is to optimize image denoising or image super resolution to maximize the success of image retrieval by using the quality index we define in Section 3. Since optimizing arbitrary denoising or super resolution methods for such an index appears difficult, we present a framework where both denoising and super resolution are achieved through CNNs. We believe that this is a reasonable approach since deep CNN based methods have also been shown to achieve state of the art enhancement results. Further, the use of a CNN to define the quality with respect to image retrieval allows for a differentiable cost function. Thus gradients can be computed during back propagation to update the denoising or super resolution CNN. Note that while the true average precision can be computed for every distorted image, it is not clear how to write a differentiable cost function that can be used to update the enhancement CNN. Computing the true average precision involves finding matching scores for every image in the database with respect to the query image and listing out images from the database based on a threshold on the matching scores. Further, the threshold needs to be varied to obtain the average precision. Our CNN based approach allows to predict the average precision using a differentiable cost function. We first present the details of image denoising in detail. Super resolution follows similarly.

4.1. Image denoising for image retrieval

The proposed architecture is given in Fig. 5. As shown in the figure, there are two CNNs, one for image denoising and another for QAIR. Initially, both these CNNs are trained independently (details mentioned in Section 6.2). Then, during the combined training stage, the QAIR CNN weights are kept frozen. Only the denoising CNN is fine tuned, by minimizing a combined loss function based on the output of the QAIR CNN and the mean squared error (MSE) of the denoised image with respect to the reference image. In particular, the combined loss function is defined as

$$L = (1 - AP) + \lambda * MSE, \tag{3}$$

where AP is the average precision predicted by the QAIR CNN and MSE is the mean square error between the denoised image and the reference image. This combined loss function ensures that the quality

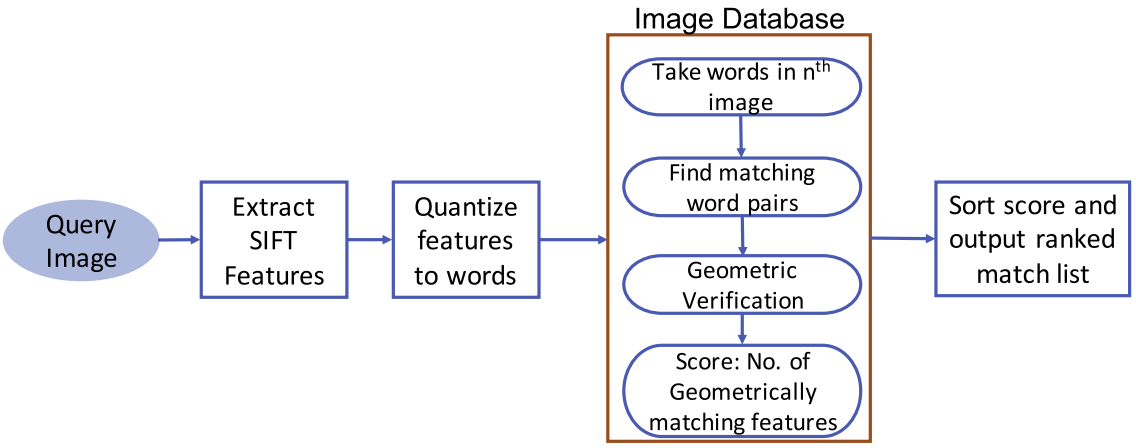

Fig. 1. A block diagram of the steps in an image retrieval algorithm.

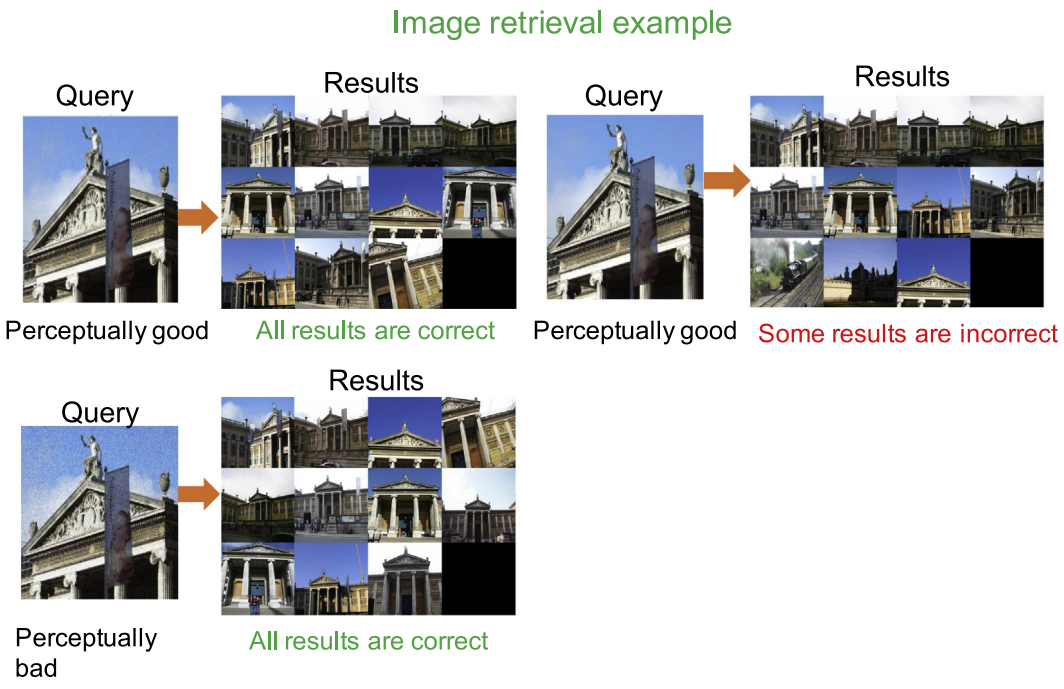

Fig. 2. Difference between perceptual quality and quality assessment for image retrieval.

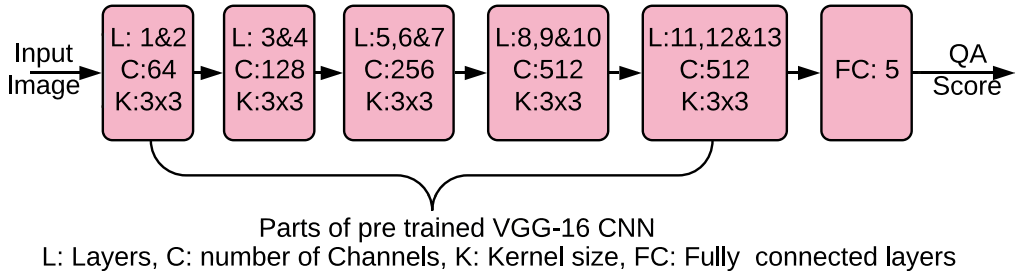

Fig. 3. Block diagram of QAIR CNN.

as predicted by the QA CNN improves without changing the denoised image too much from the actual image. λ is a parameter used to balance the two losses and the optimal value is learnt through a validation dataset. Note that in the combined loss function, AP is a function of weights of both the denoising CNN and the QA CNN, whereas MSE is a function only of the former. Once the combined training is over, the denoising CNN alone can be used for denoising and testing.

While several CNN architectures have been proposed in literature for denoising [9,20], we use a deep CNN based on the work by Zhang et al. [9], which predicts the residual noise in a noisy image. This residual noise image when subtracted from the noisy image gives the clean image. A block diagram of the network is given in Fig. 6. This CNN has 20 layers and each layer has 64 channels. All convolution kernels are of size 3×3.

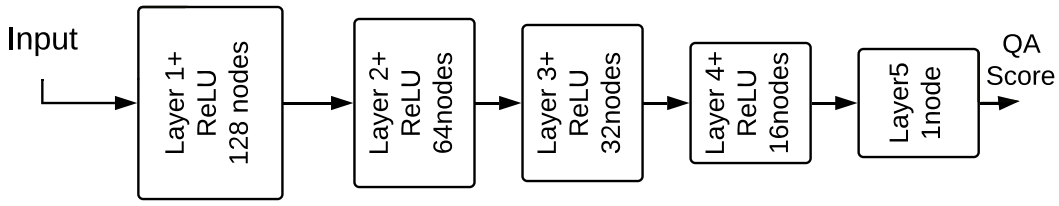

Fig. 4. Details of newly added fully connected layers.

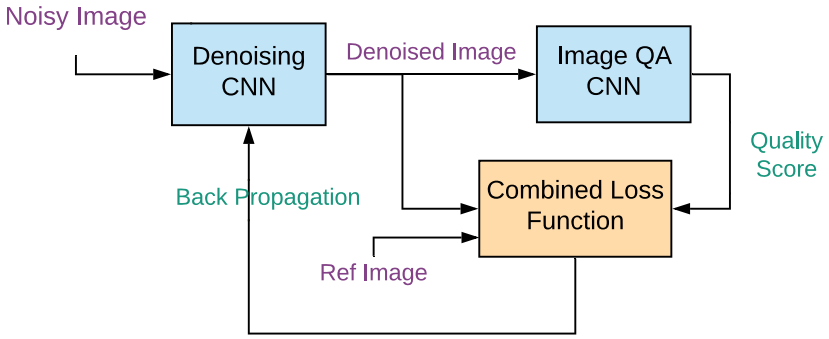

Fig. 5. Training phase of the denoising network with QAIR CNN.

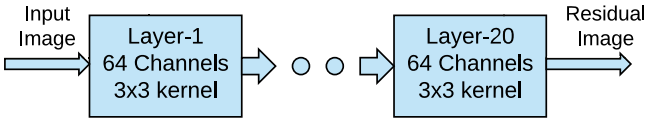

Fig. 6. Block diagram of denoising CNN [9].

4.2. Image super resolution for image retrieval

In addition to image denoising, we also consider the task of image super resolution for image retrieval. The framework we adopt is very similar to the above for image denoising, where the denoising CNN is replaced by the super resolution CNN. While several algorithms for single image super resolution exist in literature, we focus on CNN based approaches and optimize the CNNs for image retrieval using our QAIR network. In particular, we employ the DnCNN [9] used above for denoising, since it also achieves state of the art performance for image super resolution [9]. Given an up-sampled image using bi-cubic interpolation, the CNN is trained to predict the residual image, or the difference between the reference image and the upsampled image. The residual image is then added to the bi-cubic interpolated image to obtain the super resolved image. This CNN has 20 layers and each layer has 64 channels. All convolutional kernels are of size 3×3.

5. Feature adaptation for image retrieval

So far, we explored image enhancement for image retrieval. However, since retrieval is primarily based on image features, we now explore the problem of feature adaptation for image retrieval. The intuition behind this approach is that since features are ultimately used for retrieval, one could potentially perform better by adapting the features to account for distortions in addition to enhancing the images. We address this question in the context of a deep CNN based image retrieval algorithm [11], since it allows the flexibility to modify the features by changing the weights of the CNN. Thus, we consider whether we can improve retrieval performance by applying feature adaptation on top of image enhancement. Note that in Section 4, we fixed the feature vector and optimized image enhancement with respect to the given feature vector. However, we now fix the image enhancement and ask whether the feature extraction process can be optimized to improve image retrieval performance. In each of these two methods, different sets of parameters are optimized and one approach does not include the other.

5.1. Feature adaptation framework

We illustrate our framework for feature adaptation in Fig. 7. First, we pass the degraded image through an image enhancement network, potentially fine tuned as discussed in Section 4. The feature extraction procedure is then carried out on this enhanced image. The output of the feature extraction network is given as input to a feature QA CNN. We introduce the feature QA CNN to predict the performance of the features in terms of average precision as a measure of the success of the retrieval algorithm. A combined loss function based on the output of the feature QA CNN and the features of the enhanced image is used to fine tune the weights of the feature extraction network. The loss function is represented as

$$L = (1 - AP) + \lambda M, \tag{4}$$

where AP is the average precision predicted by the feature QA CNN and M is the mean square error between the present output of the feature extraction CNN and the initial output of the feature extraction CNN in its original configuration. The first term updates the feature extraction CNN such that the retrieval performance improves, while the second term ensures that the feature extraction CNN output does not deviate too much.

5.2. Feature QA network

We design the feature QA CNN to predict the average precision of the image given the output of the feature extraction CNN. The output of the feature extraction CNN, is a 4D tensor with shape $N \times 36 \times 36 \times 512$ where N is the batchsize. The feature QA CNN contains a global average pooling layer and five fully connected layers similar to Fig. 4. The global average pooling layer converts the 4D tensor to a 2D tensor. The fully connected layers are designed so that the size gradually reduces to 1 from 128. The first four fully connected layers have ReLU activation functions and the last layer has linear activation. A block diagram of the feature QA CNN is shown in Fig. 8.

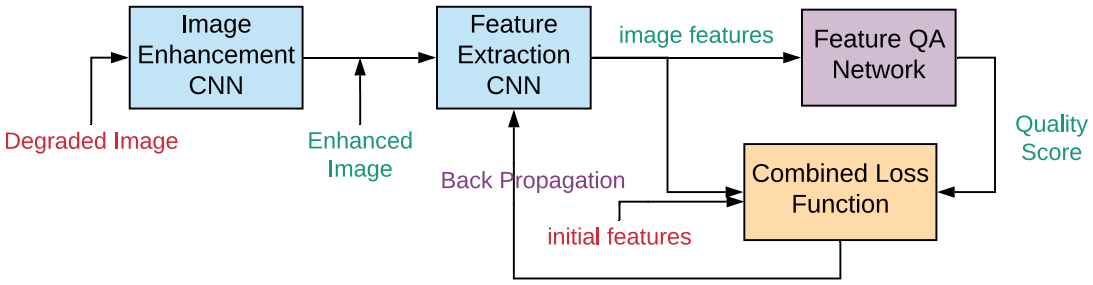

Fig. 7. Block diagram of feature enhancement framework.

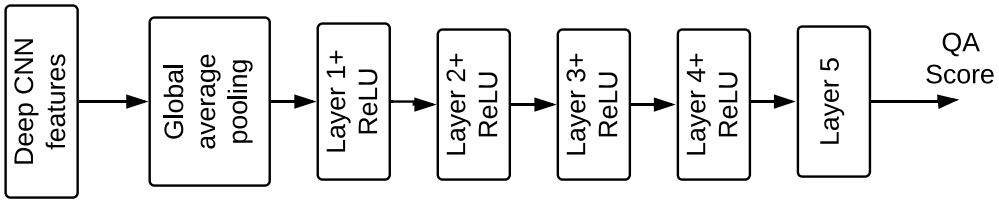

Fig. 8. Block diagram of feature QA network.

6. Experiments and results

We present the experimental results with respect to image denoising and super resolution for image retrieval and the feature adaptation framework. Before presenting the results, we describe the databases used for the experiments and the training details of the various CNNs involved. We use the SIFT based image retrieval algorithm described earlier for many of the experiments in the following unless otherwise stated.

6.1. Database

We use two standard image retrieval datasets in our experiments, the Extended Oxford buildings dataset [1] and the Paris landmarks dataset [41]. The extended Oxford buildings dataset has 5063 images of 11 different Oxford landmarks and one hundred thousand distractor images which makes a total of 105K images. Each landmark has 5 query images along with ground truth match details resulting in a total of 55 query images. The Paris dataset has 6412 images which contains 11 different Paris landmarks and 5 different query images per landmark. Out of the 55 query images in each dataset, 80%, i.e. 44 images and their distorted versions are used for training and the remaining 11 images and their distorted versions are used for testing. The train–test split is repeated across 5 iterations with a different split in each iteration such that there is no overlap in the content between the training and testing sets.

6.2. Training details

We adopt a patch based approach for image enhancement to use batch mode training. Therefore, each image is split into multiple patches of size 124 × 124. The enhancement and QAIR CNNs are trained and tested on patches. While training, all patches in an image are assigned the same quality score as the average precision on the full image. After enhancing, the image patches are merged to create the full image. All models are implemented in Python with Keras library using Tensorflow back end. We now describe the training of the CNNs used for denoising, super resolution and feature adaptation.

6.2.1. Image denoising

The image denoising CNN is trained on around 700K patches of size 50 × 50, generated from the image retrieval datasets. Different noise standard deviations are used for training the denoising CNN. We train the network for 40 epochs using Adam [42] optimizer.

During the fine tuning phase of the denoising CNN using the QAIR CNN, the images input to the QAIR CNN will be denoised images. Thus, the QAIR CNN needs to be trained on denoised images. Since the denoised images will have some amount of residual noise and blur artifacts, we create a dataset based on the 55 query images and obtain 10 different degraded versions of the same. The degradations include five different additive Gaussian noisy versions with noise standard deviation of $[3, 8, 14, 26, 44]$ and 3 blurred versions with a Gaussian blur kernel of standard deviations $[0.5, 1, 2]$. The degraded set also contains denoised images corresponding to a noise standard deviation of 50, denoised using the DnCNN [9] and BM3D algorithms [6]. The denoising CNN is fine tuned on noisy images that are split into patches of size 124 × 124. The fine tuning of the denoising CNN is performed for 40 epochs with SGD optimizer and a learning rate of 10^{-3}.

6.2.2. Image super resolution

The super resolution (SR) CNN is trained on patches of size 50 × 50. Around 700K image patches are used for training on each of the Oxford buildings dataset and the Paris dataset. The SR CNN is trained for a given super resolution factor (4 in our experiments) for 40 epochs using Adam optimizer with a mean square error loss function. The batch size is fixed to 100.

The output of the SR CNN may still have some amount of blur artefacts. We train the QAIR CNN for super resolution on varying degrees of blur to account for residual blur in the super resolved image. The degraded images include 3 blurred versions with Gaussian blur kernels of standard deviation of $[0.5, 1, 2]$ and upsampled images using bi-cubic interpolation for a scaling factor of 2 and 4. The SR CNN is fine tuned for 40 epochs using SGD optimizer with a learning rate of 10^{-3}.

6.2.3. Feature adaptation

We let the images processed by the DnCNN [9] as above for image enhancement pass through the feature adaptation CNN. In order to fine tune the feature extraction CNN, we first train the feature QA CNN on the output of the feature extraction CNN for different types of image distortions. For the denoising case, degraded versions include different levels of noisy and blurred images whereas for super resolution, the

Table 1
Mean absolute error (MAE) between predicted and actual QA scores for denoised images corresponding to noise standard deviation $\sigma = 50$.

Dataset	Oxford	Paris
MAE	0.086	0.060

Table 2
Mean absolute error (MAE) between predicted and actual QA scores for image super resolution by a factor of 4.

Dataset	Oxford	Paris
MAE	0.071	0.036

Table 3
Mean average precision for noisy and denoised images for noise standard deviation $\sigma = 50$.

Dataset	Clean	Noisy	BM3D	NN-Org	NN	NN-QA
Oxford	0.632	0.487	0.590	0.589	0.585	**0.595**
Paris	0.633	0.562	**0.615**	0.610	0.608	0.611

degradations include different levels of blurred and downsampled–upsampled images. The feature QA network is trained using Adam optimizer with mean absolute error loss function for 60 epochs.

The fine tuning of the feature extraction network is done using degraded images of similar degradation levels on which we want to test. The feature extraction CNN is trained using SGD optimizer with a low learning rate of 3×10^{-4}.

6.3. QAIR performance

We evaluate the performance of the QAIR-CNN with respect to denoising in Table 1 by computing the mean absolute error between the actual average precision and predicted average precision. We test its quality prediction performance on images denoised using the CNN for a noise standard deviation of 50. As mentioned before, the evaluations are performed by the splitting the dataset of denoised images into training and testing in the ratio 80:20 ensuring no overlap of scene content between training and testing and averaging the performance across 5 iterations. The results indicate that the mean absolute error is quite low and the QAIR CNN is able to predict the image retrieval performance reasonably well.

Further, we measure the accuracy of prediction of the retrieval performance by the QAIR CNN with respect to super resolution in Table 2. The results indicate that the mean absolute error between the predicted and actual average precision scores is reasonable.

6.4. Image denoising

We now present the results of image denoising for image retrieval for different ranges of noise levels in the following subsections.

6.4.1. Denoising for noise standard deviation of 50

We first present the results of denoising for a noise standard deviation $\sigma = 50$. The results for Oxford and Paris dataset are given in Table 3. In this table, "Clean" refers to the mean average precision on the clean images, "Noisy" refers to the same on the noisy images, "BM3D" denotes the results of images denoised using BM3D algorithm, "NN-Org" refers to the pre-trained CNN as in [9], "NN" denotes the CNN trained by us using images in the retrieval datasets and "NN-QA" refers to the results of our method i.e. the fine tuned denoising CNN using the QA CNN.

As seen in the table, our method outperforms all other methods in Oxford dataset. On the Paris dataset, the performance of our method is slightly less than that of BM3D, but better than the pre-trained CNN and the CNN trained by us. We show examples of the noisy image and denoised images using different techniques in Fig. 9. The image denoised using our technique is visually sharper than the other images and also achieves a better average precision.

Table 4
Mean average precision for noisy and denoised images for noise standard deviation $\sigma = 90$.

Dataset	Clean	Noisy	BM3D	NN-Org	NN	NN-QA
Oxford	0.632	0.158	0.455	NA	0.483	**0.499**
Paris	0.633	0.367	0.530	NA	0.556	**0.562**

Table 5
Mean average precision for noisy and denoised images for noise standard deviation in the range 30–60.

Dataset	Clean	Noisy	BM3D	NN-Org	NN	NN-QA
Oxford	0.632	0.539	0.605	0.602	0.605	**0.618**
Paris	0.633	0.589	**0.622**	0.617	0.618	0.619

6.4.2. Denoising for a noise standard deviation of 90

We evaluate the performance of our algorithm at a higher noise level of $\sigma = 90$. The denoising CNN is trained on the dataset for this noise level. The mean average precision for different methods are given in Table 4. The results show that our denoising method is superior to all other denoising methods in terms of image retrieval performance. We note that the improvements with respect to the other methods are slightly higher for $\sigma = 90$ when compared to $\sigma = 50$.

6.4.3. Denoising for variable noise levels

While so far we evaluated our algorithm on fixed noise levels, we now test our algorithm for noise standard deviations belonging to a range between 30 and 60. We do not consider noise levels corresponding to standard deviation less than 30 since there is very minimal drop in the retrieval performance at those noise levels. Here, the denoising CNN is trained on a random subset of noise standard deviations in the range of 30 to 60. The results of our method and other methods for the variable noise level case is given in Table 5.

Across the three sets of results described above, we observe that BM3D is competitive and even achieves slightly better performance than our framework sometimes on the Paris dataset. We believe that the Paris dataset has a larger set of matching blocks which lends itself to superior denoising performance of BM3D. However, we note that for the BM3D algorithm, we need to specify the exact noise standard deviation while the other methods do not require such knowledge. We observe that on the Oxford dataset, our method performs better than all the other methods and on the Paris dataset, the performance of our method is only slightly less than that of BM3D which requires knowledge of the noise standard deviation.

6.5. Image retrieval using SURF features

We now evaluate the performance of our method for another image retrieval method based on Speeded Up Robust Features (SURF) [43]. SURF is a speeded up version of SIFT and uses box filters to approximate difference of Gaussian. The feature length is 64 for SURF, in comparison to 128 for SIFT. Initially, the SURF features are computed for all images in the database and stored. A given query image is then compared with all images in the database based on the number of matching SURF features and a geometric consistency check.

We present results of this method on the Paris dataset for noise levels of standard deviation $\sigma = 50$ and $\sigma = 90$. The image retrieval results in Table 6 reveal that for $\sigma = 50$, our method performs almost as well as the original DnCNN, and better than the DnCNN trained by us. For $\sigma = 90$, our method outperforms all other methods. The performance of "NN-Org" is marked as "NA" since that CNN is not trained for this noise level.

Fig. 9. (a) Noisy image for $\sigma = 50$; (b) denoised image using BM3D; (c) denoised image using the pre-trained CNN in [9]; (d) denoised image using the CNN in [9] trained by us (e) denoised image using a CNN which is trained in combination with QA network.

Table 6
Mean average precision for noisy and denoised images for SURF based image retrieval.

σ	Clean	Noisy	BM3D	NN-Org	NN	NN-QA
50	0.392	0.251	0.355	**0.357**	0.352	0.355
90	0.392	0.057	0.222	NA	0.259	**0.267**

Table 7
Mean average precision for noisy and denoised images for CNN features based image retrieval.

Dataset	Clean	Noisy	BM3D	NN-Org	NN	NN-QA
Oxford	0.451	0.014	**0.220**	0.141	0.214	0.219
Paris	0.658	0.212	**0.513**	0.436	0.486	0.487

6.6. Image retrieval using deep CNN features

We also test our image enhancement framework on a different image retrieval method using deep features [11]. In this method, a pre-trained deep CNN, VGG-19 [12], is used to extract features from the images. The last convolutional layer output of the network is used as features. A block diagram of the feature extraction and aggregation steps for image retrieval are shown in Fig. 10. The experimental results for $\sigma = 50$ on both the Oxford and Paris datasets are shown in Table 7. While the improvements with respect to other learning methods are marginal and there is a performance gap with respect to BM3D, we show later on that the performance of our framework can be further improved using feature adaptation.

6.7. Super resolution results

We now present the performance analysis of our enhancement framework for super resolution in Table 8. In the table, "Original" refers to the actual query image, "Down sample by 4" refers to the image down sampled by 4, "Up sample bicubic 4" refers to the image which is down sampled and then bi-cubic interpolated by a factor of 4, "SR NN" refers to the images output by the SR CNN, "SR CNN QA" refers to the output of the fine tuned SR CNN using the QAIR CNN. The results indicate that our method outperforms all other methods on both the Oxford and Paris datasets. We also observe that super resolution of downsampled images can sometimes improve the retrieval performance. We believe this can be viewed as some preprocessing of the image before image retrieval that can improve retrieval performance.

An example image and its different SR versions are given in Fig. 11. We see that the image enhanced using our method is sharper than other images leading to a performance improvement in terms of average precision.

Table 8
Mean average precision for image super resolution by a factor of 4.

Dataset	Original	Down sample by 4	Up sample bi-cubic 4	SR NN	SR NN QA
Oxford	0.632	0.265	0.604	0.620	**0.626**
Paris	0.633	0.323	0.619	0.635	**0.642**

Table 9
Mean average precision for both noisy and low resolution (LR) images using CNN based enhancement and image retrieval.

Dataset	Clean	Noisy/LR	NN-Org	NN	NN-QA
Paris	0.6330	0.6020	0.6128	0.6308	**0.6322**

Table 10
Mean average precision for noisy and denoised images for noise standard deviation $\sigma = 50$.

Dataset	Clean	Noisy	BM3D	NN-Org	NN	NN + tweaked features
Oxford	0.451	0.014	0.220	0.141	0.214	**0.230**
Paris	0.658	0.212	**0.513**	0.436	0.486	0.507

6.8. Denoising and super resolution using the same network

We also conduct an experiment where both denoising and super resolution are achieved using a single enhancement CNN on the Paris dataset. The corresponding QAIR CNN is trained on a mix of denoised, noisy, blurred and low resolution images upsampled using bicubic interpolation and CNN based algorithms. Further, the enhancement CNN was trained as before on 8 distorted versions of each query image, with 4 noisy images with noise standard deviation between 30 and 60 and 4 low resolution images with downsampling factors of 3.8, 3.9, 4 or 4.1. The test set consists of 2 distorted versions of each query image with one noisy image with standard deviation between 30 and 60 and 1 low resolution image with downsampling factor of 4. The same train–test split among query images explained earlier across multiple iterations was used. The results in Table 9 indicate that the "NN-QA" approach does indeed offer benefits when compared to not using the QAIR CNN for training the enhancement CNN.

6.9. Feature adaptation

We now analyze the performance of the feature adaptation framework for both denoising and super resolution.

6.9.1. Feature adaptation for noisy images with $\sigma = 50$

The image retrieval performance on the Oxford and Paris dataset are shown in Table 10. In this table, "NN+tweaked features" refers to the results of our proposed method in which the images are passed through a denoising CNN and the feature extraction CNN is fine tuned

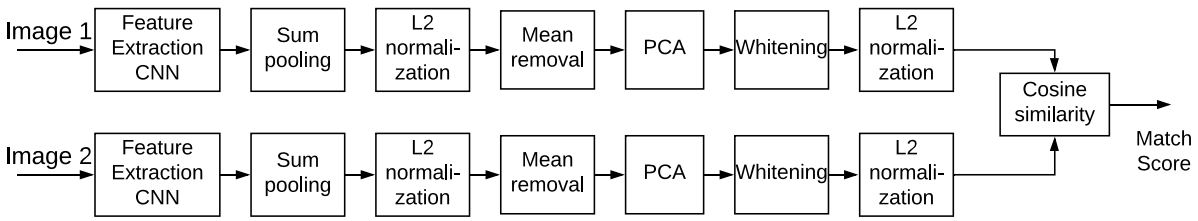

Fig. 10. Block diagram of CNN based feature extraction and image matching.

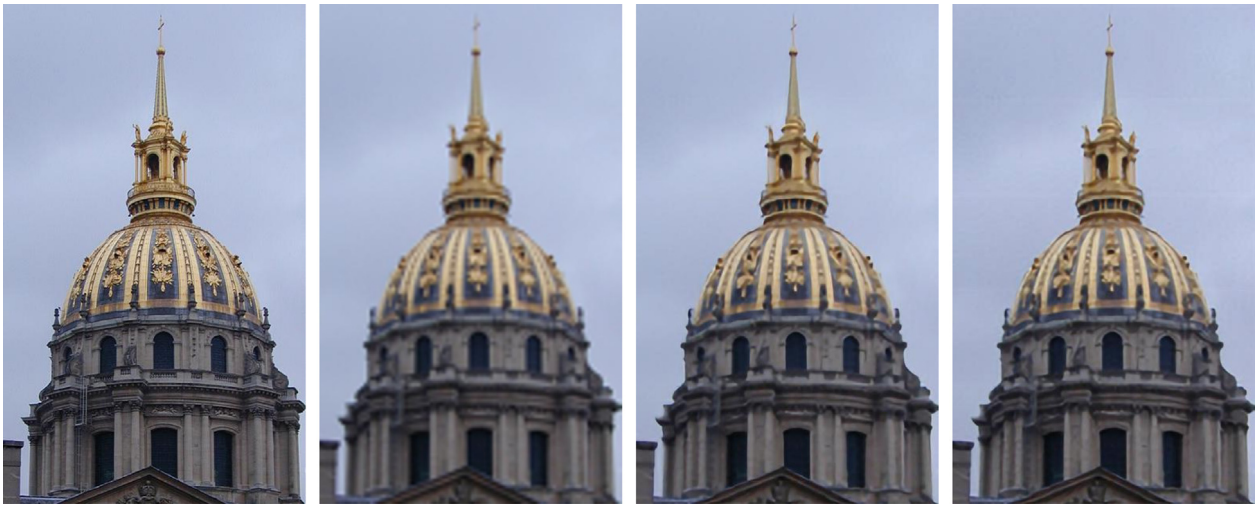

Fig. 11. (a) Original Image; (b) Image interpolated using bicubic interpolation; (c) super resolution image using the CNN in [9] trained by us (d) super resolution image using a CNN which is trained in combination with QA network.

Table 11
Mean average precision for noisy and denoised images for noise standard deviation in the range 30 to 60.

Dataset	Clean	Noisy	BM3D	NN-Org	NN	NN + tweaked features
Oxford	0.451	0.047	0.235	0.156	0.222	**0.255**
Paris	0.658	0.276	**0.536**	0.468	0.513	0.533

using a feature QA CNN. The results shows that fine tuning the feature extraction CNN can lead to a performance improvement. We also note that the feature enhancement method performs better when compared to image enhancement in Table 7 for the same scenario.

6.9.2. Feature adaptation for noisy images with σ in the range $[30, 60]$

We now analyze how the feature adaption method performs when we train for noise levels in a range instead of a single noise level. Here, we train for noise values with standard deviation in the range $[30, 60]$. The feature QA network is trained in a similar setup as that of the previous section. The results in Table 11 indicate that there is a good improvement in both datasets for our method when compared to the case where features are not fine tuned.

6.9.3. Feature adaptation results for image super resolution by a factor of 4

We also test the feature adaptation framework for the super resolution case. We design the feature adaptation framework by first passing the low resolution image though a super resolution network and then through a fine tuned feature extraction network. The image SR CNN and its training method are same as that in Section 4.2. The feature QA network is trained on query images and different degraded versions of the same as discussed before. The results obtained on the Oxford and Paris dataset are given in Table 12. In the table, "SR NN + tweaked features" refers to the performance of SR CNN output using enhanced features. Again, the results show that our method out performs all other methods.

In summary, we observe that our image enhancement and feature adaptation frameworks yield improvements in the mean average precision in several scenarios. In other scenarios, the performance is almost as good as the best performing enhancement method.

7. Conclusion and future work

In this work, we developed a framework for image enhancement for image retrieval by defining a relevant notion of image quality. The quality of a query image for image retrieval is defined as the area under the precision–recall curve for that image and we designed a deep CNN based method for image quality prediction. By modeling the image enhancement as a deep CNN, we can fine tune such a network for the success of image retrieval. Note that this particular modeling is not very restrictive owing to the success of deep CNN methods in a variety of image processing applications. We showed the benefits of such an approach for two image enhancement cases, image denoising and image super resolution.

We also developed a framework for feature adaptation to improve the image retrieval performance of degraded images using a feature QA network. This framework of feature adaptation is applicable for

Table 12
Mean average precision for image super resolution by a factor of 4.

Dataset	Original	Down sample by 4	Up sample bi-cubic 4	SR NN	SR NN + tweaked features
Oxford	0.451	0.015	0.150	0.166	**0.186**
Paris	0.658	0.296	0.397	0.496	**0.537**

deep CNN features based image retrieval. We tested such an approach for both image denoising and image super resolution cases and were successful in showing that fine tuning the feature extraction framework using a feature QA network leads to better image retrieval performance.

While we showed the utility of our framework, one could potentially improve the results by further enhancing the prediction of average precision. Moreover, we considered the homogenous distortions and a patch based approach for enhancement. It will be interesting to study enhancement operations that operate on the entire image with the average precision of the entire image. In order to attempt such an approach, methods for predicting the average precision using much lesser data may need to be explored. Further, the framework can be extended to study other enhancement settings such as low light enhancement, defogging and so on.

While we considered the specific case of image retrieval, our framework can also be used to study image enhancement for various other computer vision applications. Our framework is particularly useful when the performance of the computer vision task is arbitrary and cannot be modeled by closed form expressions of the output of a deep CNN. Thus, by converting any computer vision task performance to a quality assessment CNN, one could potentially optimize image enhancement for the relevant computer vision task.

Acknowledgment

This research was supported by a grant from Robert Bosch Center for Cyber Physical Systems (RBCCPS), Indian Institute of Science.

References

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[2] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: Improving particular object retrieval in large scale image databases, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[3] J. Philbin, M. Isard, J. Sivic, A. Zisserman, Descriptor learning for efficient retrieval, in: European Conference on Computer Vision, 2010, pp. 677-691.

[4] R. Arandjelović, A. Zisserman, Three things everyone should know to improve object retrieval, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2911-2918.

[5] H. Jégou, M. Douze, C. Schmid, Improving bag-of-features for large scale image search, Int. J. Comput. Vis. 87 (3) (2010) 316–336.

[6] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Image denoising by sparse 3-D transform-domain collaborative filtering, IEEE Trans. Image Process. 16 (8) (2007) 2080–2095.

[7] D. Zoran, Y. Weiss, From learning models of natural image patches to whole image restoration, in: International Conference on Computer Vision, Nov 2011, pp. 479-486.

[8] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.

[9] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.

[10] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[11] A.B. Yandex, V. Lempitsky, Aggregating local deep features for image retrieval, in: 2015 IEEE International Conference on Computer Vision (ICCV), Dec 2015, pp. 1269-1277.

[12] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556.

[13] B.V. Somasundaran, R. Soundararajan, S. Biswas, Image denoising for image retrieval by cascading a deep quality assessment network, in: Proceedings of IEEE International Conference on Image Processing (ICIP), Oct 2018, ser. ICIP '18, 2018.

[14] Y. Cao, C. Wang, Z. Li, L. Zhang, L. Zhang, Spatial-bag-of-features, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3352-3359.

[15] Y. Zhang, Z. Jia, T. Chen, Image retrieval with geometry-preserving visual phrases, in CVPR 2011, 2011, pp. 809-816.

[16] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2011) 1704–1716.

[17] J. Portilla, V. Strela, M.J. Wainwright, E.P. Simoncelli, Image denoising using scale mixtures of Gaussians in the wavelet domain, IEEE Trans. Image Process. 12 (2003) 1338–1351.

[18] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (12) (2006) 3736–3745.

[19] H.C. Burger, C.J. Schuler, S. Harmeling, Image denoising: Can plain neural networks compete with bm3d? in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 2392-2399.

[20] K. Zhang, W. Zuo, L. Zhang, FFDNet: Toward a fast and flexible solution for CNN based image denoising, CoRR, 2017. [Online]. Available: http://arxiv.org/abs/1710.04026.

[21] M. Irani, S. Peleg, Improving resolution by image registration, CVGIP: Graph. Models Image Process. 53 (3) (1991) 231–239, [Online]. Available: http://dx.doi.org/10.1016/1049-9652(91)90045-L.

[22] W.T. Freeman, T.R. Jones, E.C. Pasztor, Example-based super-resolution, IEEE Comput. Graph. Appl. 22 (2002) 56–65.

[23] D. Glasner, S. Bagon, M. Irani, Super-resolution from a single image, in: 2009 IEEE 12th International Conference on Computer Vision, Sept 2009, pp. 349-356.

[24] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, CoRR, vol. abs/1501.00092, 2015. [Online]. Available: http://arxiv.org/abs/1501.00092.

[25] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 1646-1654.

[26] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the dct domain, IEEE Trans. Image Process. 21 (8) (2012) 3339–3352.

[27] M.A. Saad, A.C. Bovik, C. Charrier, Blind prediction of natural video quality, IEEE Trans. Image Process. 23 (3) (2014) 1352–1365.

[28] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.

[29] P. Ye, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 1098-1105.

[30] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, in: CVPR '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 1733–1740, [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.224.

[31] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, CoRR, vol. abs/1602.05531, 2016. [Online]. Available: http://arxiv.org/abs/1602.05531.

[32] K. Ma, W. Liu, T. Liu, Z. Wang, D. Tao, DipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs, IEEE Trans. Image Process. 26 (8) (2017) 3951–3964.

[33] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-end blind image quality assessment using deep neural networks, IEEE Trans. Image Process. 27 (3) (2017) 1202–1213.

[34] S. Bosse, D. Maniry, K. Muller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Trans. Image Process. 27 (1) (2018) 206–219.

[35] S. Gunasekar, J. Ghosh, A.C. Bovik, Face detection on distorted images augmented by perceptual quality-aware features, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2119–2131.

[36] R. Soundararajan, S. Biswas, Machine vision quality assessment for robust face detection, Signal Process., Image Commun. 72 (2019) 92–104.

[37] D.M. Rouse, R. Pépion, S.S. Hemami, P.L. Callet, Image utility assessment and a relationship with image quality assessment, in: Human Vision and Electronic Imaging, 2009.

[38] A. Gala, S. Shah, Joint modeling of algorithm behavior and image quality for algorithm performance prediction, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2010, pp. 31.1–31.11, http://dx.doi.org/10.5244/C.24.31.

[39] D. Liu, B. Wen, X. Liu, Z. Wang, T. Huang, When image denoising meets high-level vision tasks: A deep learning approach, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 842–848, [Online]. Available: https://doi.org/10.24963/ijcai.2018/117.

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. (IJCV) 115 (3) (2015) 211–252.

[41] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: Improving particular object retrieval in large scale image databases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[42] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[43] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Underst. 110 (3) (2008) 346–359, [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2007.09.014.