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

Convolutional Neural Networks for Relevance Feedback in Content Based Image Retrieval

A Content Based Image Retrieval System that Exploits Convolutional Neural Networks both for Feature Extraction and for Relevance Feedback

Lorenzo Putzu · Luca Piras · Giorgio Giacinto

Received: date / Accepted: date

Abstract Given the great success of Convolutional Neural Network (CNN) for image representation and classification tasks, we argue that Content-Based Image Retrieval (CBIR) systems could also leverage on CNN capabilities, mainly when Relevance Feedback (RF) mechanisms are employed. On the one hand, to improve the performances of CBIRs, that are strictly related to the effectiveness of the descriptors used to represent an image, as they aim at providing the user with images similar to an initial query image. On the other hand, to reduce the semantic gap between the similarity perceived by the user and the similarity computed by the machine, by exploiting an RF mechanism where the user labels the returned images as being relevant or not concerning her interests. Consequently, in this work, we propose a CBIR system based on transfer learning from a CNN trained on a vast image database (ImageNet [39]), thus exploiting the generic image representation that it has already learned. Then, the pre-trained CNN is also fine-tuned exploiting the RF supplied by the user to reduce the semantic gap. In particular, after the user's feedback, we propose to tune and then re-train the CNN according to the labelled set of relevant and non-relevant images. Then, we suggest different strategies to exploit the updated CNN for returning a novel set of images that are expected to be relevant to the user's needs. Experimental results on different data sets show the effectiveness of the proposed mechanisms in improving the representation power of the CNN with respect to the user concept of image similarity. Moreover, the pros and cons of the different approaches can be clearly pointed out, thus providing clear guidelines for the implementation in production environments.

Lorenzo Putzu · Luca Piras · Giorgio Giacinto
Dept. of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
E-mail: lorenzo.putzu@unica.it, luca.piras@unica.it, giacinto@unica.it
WWW home page: `http://pralab.diee.unica.it`

Luca Piras · Giorgio Giacinto
Pluribus One
Via Bellini 9, 09128 Cagliari, Italy
WWW home page: `https://pluribus-one.it`

Keywords Content Based Image Retrieval, Convolutional Neural Network, Feature extraction, Similarity, Relevance feedback

1 Introduction

Description, recognition, and automatic classification of the structures in the images are the basis of a large number of applications that require the processing and transmission of visual information. These applications are based on image processing techniques aimed to extract information that is tailored to the task at hand. The information extracted from the images is then analysed to provide visual or logical patterns based on the characteristics of the images and their mutual relationship. Content-Based Image Retrieval (CBIR) is one of such applications that leverages on the description and representation of the information in images.

CBIR systems refer to the approaches to retrieve digital images from large databases, by analysing their visual and semantic content. The goal of these systems is to retrieve a set of images that is best suited to the user's intention that is formulated using a query image. To retrieve these images, the CBIR uses a set of distance functions to estimate the similarity between the query image and the other images in the repository applied on the features space used to describe the image content. Given their ability to create the representation of the input images internally as a result of the learning process, Convolutional Neural Networks (CNN) [4] are considered as one of the most effective techniques to extract meaningful features to describe the image content. Indeed CNN is nowadays the state-of-the-art technique in Image Classification and Image Retrieval problems, able to achieve results never achieved before [18] and beat humans in many challenges [39]. However, there is still a gap between the human perception and the description of images based on features, since the semantic content present in an image is highly subjective, and, therefore, very difficult to describe analytically. This difference in perception is known as the *semantic gap*.

Different mechanisms can be employed to fill the gap to improve the effectiveness of CBIR systems. Relevance Feedback (RF) is one of these mechanisms, that involves the user in iteratively refining the search results [36] [38]. After the user submits a query image, she can give to the system her feedback to the search results by labelling the returned images as relevant to the query or not. The system, then, uses this information to provide a more significant number of related images in the next iteration, by reformulating the original query, by re-weighting the features according to the relevant images or by estimating the posterior probability distribution of a random variable according to the user feedback. A different group of approaches use a formulation of RF in terms of a *pattern classification* task, by using the relevant and non-relevant image sets to train popular learning algorithms such as SVMs [20], neural networks and self-organising maps [27,32]. Even CNNs can be re-purposed for this task as proposed in [47,49,31,48] to move the non-relevant image away from the query image through a modification of the feature space using the back-propagation algorithm.

Based on that idea, here we investigated the use of CNNs both for image representation and RF, evaluating two different architectures both to extract new features or to classify the images according to the relevance of their contents. According to our investigation, we recommend the use of CNN for feature extraction

in image retrieval tasks as is, without a fine-tuning process, as already proposed in [35,12], where they highlighted the effectiveness and generality of the learned CNN representations. But mostly, we would like to discourage the CNN fine-tuning procedure for image retrieval tasks, which has recently caught on [49,48], where they tuned the CNNs by using the whole original data sets. Indeed, while in image classification tasks this approach is already widely used and generally recognised, we argue that in CBIR tasks this approach goes against the idea of retrieval in which there is no concept of classes (see details in Sect. 3).

We propose two CNN architectures to exploit relevance feedback, both derived from the well know AlexNet [18] model. The first one preserves the original network depth, and it has been adapted to the RF task by modifying just the last layer, while the second one has been adapted by adding a further layer. Both networks are fine-tuned to separate the relevant and non-relevant and re-trained using the images labelled by the user as the training set. Since this training set is often tiny and unbalanced, we suggested an approach to creating a richer and more representative training set.

The main contributions of this paper are the proposal of two strategies to create a new image ranking, one strategy based on the exploitation of the CNN as a feature extractor, the other strategy based on the use of CNN as a classifier, where the two classes are "relevant" and "non-relevant". The first strategy allowed us to propose three different methods to refine the query according to the user feedback either by refining or reformulating the query that could be misplaced or marginal to the user's intention with respect to the relevant images in the database.

The remainder of the manuscript is structured as follows: CBIR systems are presented in Section 2, detailing the features used for image retrieval and the use of an RF phase. Section 3 presents out CBIR system based on generic CNN features. The CNN architectures and strategy to exploit RF are described in Section 4 with details on the creation of the training set. In Section 5, we introduce the materials and methods used in our experiments, that are presented in detail in Section 6. Finally, Section 7 is devoted to the query refinement procedures, the analysis and discussion of the overall results and Section 8 to conclude.

2 Background

Many CBIR systems, based on the automatic analysis of the image content from a computer perspective, have been proposed over the years. Most of these systems are designed to tackle specific applications and consequently to specific retrieval problems. An example comes from the Computer-Aided Diagnosis systems [16], that can help the diagnosis by performing queries on a large database of images already labelled by specialists. Other examples are related to sport events [28], cultural heritage preservation [53], identification of products for comparative shopping [25]. As it can be guessed, a CBIR designed for a specific task can be easily managed and can obtain more satisfactory results than a non-specific CBIR that works on large and generic databases [46]. For this reason, the design of a general-purpose multimedia retrieval system is still a challenging task, as such a system should be capable of adapting to different semantic contents, and different intents of the users.

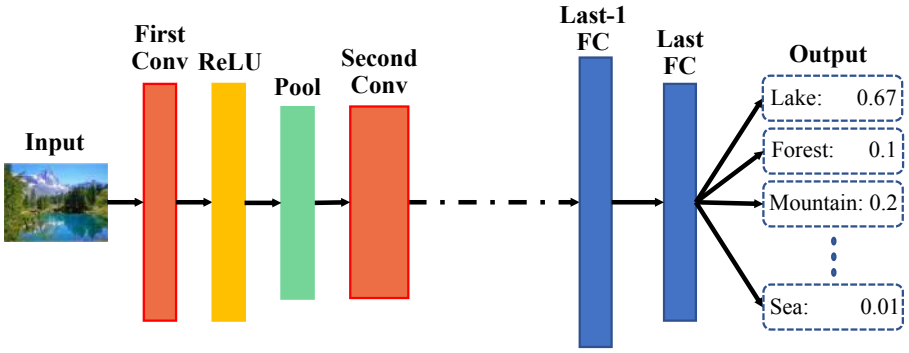

Fig. 1 An example of CNN architecture.

2.1 Features in CBIR

In CBIR systems, the selection and the representation of a set of content-based features, that are expected to capture the semantics of the images, are essential steps. Indeed, the accuracy in retrieval depends heavily from a robust set of features. While some CBIR systems employ low-level and mid-level features [10], trying to take into account information like colour, edge and texture [1], CBIR systems that address specific retrieval problems leverage on different kind of features that are specifically designed [45]. Some works exploited low-level image descriptors, such as colour histograms [13], a fusion of textual and visual information [17,44, 41] or even Scale Invariant Feature Transform (SIFT) [52,51] originally proposed for object recognition [23] and then extended to scene categorisation by using the Bag-of-Visual-Word (BoVW) model [19]. Even more specific low-level features designed for other applications have been used in CBIR systems, such as the HOG [8] and the LBP [29] descriptors, originally proposed for pedestrian detection and texture analysis respectively. Several CBIR systems use a combination or fusion of different image descriptors [5] to provide a high-level understanding of the scene. A simple solution is based on the concatenation of the feature vectors, such as in [52], where the authors propose two ways based on BoVW of integrating the SIFT and LBP. A different solution produces a fusion of the output by combining either different similarities or distances from the query [14] or different ranks obtained by the classifiers[40].

However, it is not easy to determine which features could describe the images adequately. For this reason, CNNs are popular for image classification and retrieval tasks since they can learn features by creating their representation of the input images internally as a result of the learning process [4]. The goal of the architecture of a CNN is to model high-level features by employing a high number of connected layers composed of multiple non-linear transformation units (see Figure 1). The convolutional layers are the core of a CNN as they perform most of the computations, because they convolve images with different filters, and produce the activations as responses to those filters. Intuitively, the network learns which filters are activated when some visual feature or pattern is presented in the input. Typically two different layer types follow the convolutional layer: the Rectified Linear Unit (ReLU) and pooling layers. The ReLU layer applies an element-wise activation function, such as the thresholding at zero, while the pooling layer per-

forms a down-sampling operation along the spatial dimensions since it takes just the maximum value in a region. The last layers are common fully connected layers, where each neuron of a layer is fully connected to the neurons of the previous and next layers. The activations extracted from the upper CNN layers are also excellent descriptors for image retrieval [18]. It implies that a CNN trained for a specific task has acquired a generic representation of objects that is useful for all sorts of visual recognition tasks. As a consequence, CNNs attracted the interest of many researchers in this field, that deal with semantic visual content analysis, description and retrieval (see details in Sect. 3).

2.2 Relevance Feedback

In image retrieval, the query image is used as an example to find all the images in a data set that are relevant to that query. Defining which image is relevant or not relevant to a query is not a trivial issue, in particular, if the problem must be addressed using just one image as a query. Indeed, there is still a gap between the human perception of the semantic information present in an image, and its computer description, that is typically able to identify just a small subset of semantic concepts. Moreover, the user that performs the query could not have a specific target in mind, or he could start the search with an image that is only partially related to the content that he has in mind. In both cases, the retrieval process could not be accomplished in just one step. The mechanism of RF has been developed to involve the user also in further phases of the process, in particular, to verify if the search results are relevant or not.

There are three main types of feedback. The most used is the explicit feedback, through which the user can explicitly give to the system her feedback by labelling the returned images as relevant to the query or not. The user can provide her feedback also implicitly, where the system infers the user's feedback automatically from its behaviour, such as the selected images for viewing or the duration of time spent viewing a picture. Considering that user feedback is not always available, the third type of feedback called Pseudo RF or Blind RF, that automates the manual part of RF, is widely used. In this way, the users can benefit from an improved image retrieval performance without an extensive and time-consuming interaction.

Different approaches have been proposed to exploit the feedback for refining the parameters of the search. Such as by computing a new query vector [36], or by modifying the similarity measure in such a way that relevant images have a high similarity value [33], or trying to separate relevant and non-relevant images using pattern classification techniques such as Support Vector Machines [20], Decision Trees [24], Clustering [9], Random Forests [6] or Convolutional Neural Networks [48]. Our approach processes the user's feedback and separates relevant from non-relevant images by exploiting a CNN pre-trained on a large data set as in [49,31]. Here, we make an extensive investigation on different approaches to exploit the RF when a modified CNN is used both for feature extraction and for classifying images as relevant or non-relevant to the given query.

3 CNN Features

It has been recently reported that CNNs outperformed many state-of-the-art approaches in many tasks, such as object category recognition and image classification [18], handwritten digits and character recognition [3], pedestrian detection [42] medical image applications [16]. Also, as stated before, features extracted from the upper layers of the CNN can serve as image descriptors [18], implying that a CNN trained for a specific task has acquired a generic representation of objects that is useful for all sorts of visual recognition tasks [35, 12]. This reason has dramatically facilitated the use of CNNs, especially when large amounts of data and resources are not available. Indeed, the training of CNN requires a considerable quantity of data and resources for a typical problem of image classification.

3.1 Transfer learning or fine-tuning?

Transfer learning and fine-tuning are a common alternative to training a CNN from scratch, by using a pre-trained CNN model, that could be either fine-tuned to a new classification problem or used to extract the features for a new task directly. But, since CNNs have been created for classification tasks, they must be adapted to extract the activation values as a response to a given image.

The most common approach to adapt a pre-trained CNN for feature extraction, especially if used in classification tasks (eg. CNN features to power an SVM), is to adjust the final CNN layer (by modifying the type and number of categories) and re-train the network to update its weight with the new data.

This approach is mainly devoted to improving the performances of CNN features that, after the re-training procedure, are specifically tailored to the type and number of categories of the new data. Recently this approach has also been used in image retrieval tasks [47–49], by extracting the features from a CNN re-trained using the whole data set used in the retrieval step. Nevertheless, we believe that for image retrieval tasks, this type of approach is not feasible, as the images are not grouped by categories or labels but just using a similarity concept. Furthermore, the use of features extracted from a CNN tailored for a specific task and then re-purposed to a novel generic task, has been extensively explored [35, 12], demonstrating that they have sufficient representational power and generalisation ability to perform different visual recognition tasks. Therefore, the feature extraction step of our CBIR system is performed from the original network, just exploiting the CNN internal representation already learned from a generic data set.

3.2 CNN activation

Each layer of a CNN could be used to produce a response or activation to an input image, but generally, just a few levels are suitable for feature extraction purposes. Indeed the first network layers are able to describe just some images characteristics, such as points and edges, which are then processed by the innermost layers of the network to capture high-level features. Such high-level features are suited to object recognition tasks because they include all the primitive characteristics to create a

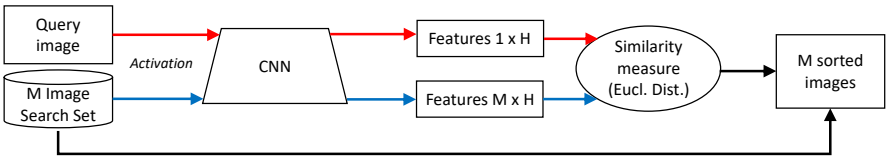

Fig. 2 Steps performed by an image retrieval system employing CNN features extracted from a generic layer. M is the number of images in the database.

strong image representation. Generally, the layer that is used for feature extraction purposes is the one that precedes the classification layer.

Thus CNN features, in an image retrieval task, are used in the same way as hand-crafted features, by defining a matrix of features, that represents the set of images to retrieve (*search set*). Consequently, if the image archive contains M images, a matrix of features of size $M \times h$, where h is the number of features (equal to the layer size), is created. Accordingly, the retrieval task can be performed by extracting the feature vector of size $1 \times h$ from the query image and computing its similarity with all the images in the archive. Then, images are sorted according to the selected similarity measure and returned to the user, as shown in Figure 2.

4 Relevance Feedback to refine CNN

In CBIR systems that make use of an RF phase, the user is prompted to give her feedback on the last retrieval phases by marking each image as being relevant or not to her query. Relevance information is then used to refine the query and provide much more relevant results in the next retrieval round.

4.1 The RFNet architectures

CNNs can be re-purposed also for the RF task by using the user's feedback to separate relevant from non-relevant images. In particular, we show how a modified CNN can produce image representations that better fit the user's needs. Since the training of a CNN is a long and computationally expensive process, we rely on a pre-trained CNN model. In this case, the network needs a fine-tuning phase to adjust the architecture and let it adapt to the type and number of categories of the new task, that is to separate the relevant images from the non-relevant ones. We used two different approaches to fine-tune CNN. The first approach consists of *replacing* the last fully connected layer with a new two-outputs layer to label images as being relevant or non-relevant. The goal of this approach is to preserve almost all the original network layers while creating a new CNN with the same depth of the original one (see Figure 3) which we refer to as *L8_RFNet*. The second approach consists in *adding* a new fully-connected two-outputs layer. In this way, the new CNN, which we refer to as *L9_RFNet* (see Figure 3), is more in-depth than the original architecture, but we can preserve all the original layers, including the weights and biases. It is worth to note that the task that we aim to perform is the update of a pre-trained network through a new training set containing a comparably tiny set of images, that is, the feedback provided by the

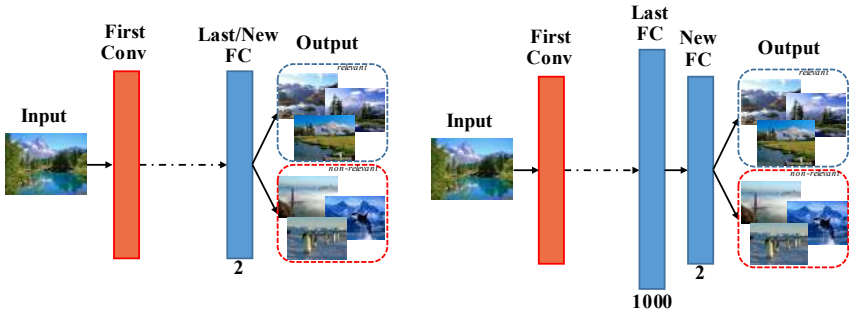

Fig. 3 Diagram of the last layers of the tuned nets used for RF: at the left L8_RFNet and at the right L9_RFNet.

user. To avoid over-fitting, and preserve as far as possible the weights and biases of the original CNN model, we have frozen the firsts network layers. Indeed, the re-training phase is just needed to create the weights and biases for the new fully connected layer and to update the information of the originals fully connected layers, with a specific image representation.

4.2 Training Set Creation from Unbalanced Set

To conclude the fine-tuning process, the RFNet is re-trained with the images belonging to the new training set. The ability to re-train a CNN that is focused on the semantic content of the query image heavily depends on the quality of the training set originated by the RF round, that is, on the relevant and non-relevant images labelled by the user. However, since the number of images labelled by the user is tiny, in particular, if compared to the ImageNet data set images [11], in most cases, the training set can be strongly unbalanced [49]. Indeed, it may occur that the training set does not contain any non-relevant images or, even worse, that it does not contain any relevant images. While in the first case the user might decide not to engage in relevance feedback as the retrieval performances are already very high, in the case in which the training set contains only non-relevant images, it means that there are no relevant images examples for training the network, except the query image.

To overcome these and other issues, in this work we defined a new and potentially richer training set, that preserves the user feedback at each RF round, and that always contains both relevant and non-relevant images. More precisely, if we set a retrieval window of size k, at the end of the retrieval round or round 0, the training set is a set of k_0 images composed of r relevant images and $k - r$ images that are not relevant. The images belonging to the $k - r$ set will no longer be shown to the user in the following rounds and therefore at the end of the first feedback round the training set will be a set of images k_1 equal to $k + (k - r)_0$. Thus, while previous feedback rounds are stored in the system, the user always has to label the same number of images, equal to k. This procedure is repeated for each feedback round, therefore during round n the training set will be equal to $k_n = k + (k - r)_{n-1}$, where the number n (with $n = 1, \ldots, 4$) in subscript represents the feedback round. As a consequence, the training set increases at each feedback round, but the user's effort remains unchanged. Moreover, to avoid a

training set composed of non-relevant images only, the query image is added to the training set, so that the training set after round 0 is a set of images of size $k + 1$, thus always having $r >= 1$.

Then, to generate more image examples from such limited image number and to prevent overfitting, we performed a simple data augmentation procedure [18, 43]. The augmented set is created by selecting eight crops of the original image (including the full image). Each crop is then horizontally flipped, blurred by an average filter and modified in colours by stretching the histograms. Hence, this data augmentation produces an augmented set of 36 images for each original image.

4.3 Proposed Relevance Feedback approaches

Once the RFNet has been tuned for a specific query according to the user's feedback, it can be used to extract the new image representation to perform a new retrieval step that benefits from RF. We propose to extract the information from the net in different ways. In particular, we focus on two main strategies: RF with CNN features extraction, and RF with classification.

In the first strategy, that exploits CNN features for RF, the RFNet is used to extract activations from the second fully connected layer (indicated as $fc7$ in Figure 6), and the similarity between images is computed according to the Euclidean distance, as shown in Figure 4. This strategy aims to exploit the new CNN's internal image representation (obtained after the re-trainining phase) to extract new features that are tailored to the query image. That is to say; relevant images features should be much closer to the query image features while non-relevant images features should be moved further away.

In the second strategy, that exploits CNN classification for RF, the RFNet is used to classify the images belonging to the search set directly. The main goal here is to create a CNN that is able to separate the relevant from the non-relevant images. Also, this strategy aims to exploit all the CNN weights and biases entirely until the output layer, which provide a further vector that contains the scores. Each score indicates the probability for an image being relevant or non-relevant as shown in Figure 4. As it can be observed, in both approaches the RF can be performed for n iterations, but the CNN architecture is tuned for the new task only once, just after round (0), while for the next n RF rounds it is only necessary to re-train the network.

4.4 Query Refinement for RF Based on CNN Features Approach

Although the conceptual representation of the two approaches looks very similar, they are very different. In particular, with the classification strategy, the query image is used just once, and the further rounds are mainly results of the user feedback, while with the feature extraction strategy the layer activations are always compared to the query image activation. That is to say that, with the feature extraction strategy, the quality of retrieval results is always related to the quality of the query image, that could be misplaced or just marginal to the concepts searched by the user. Thus, to further exploit the user feedback, the query can be reformulated through resorting to *query shifting* or *query expansion* paradigms.

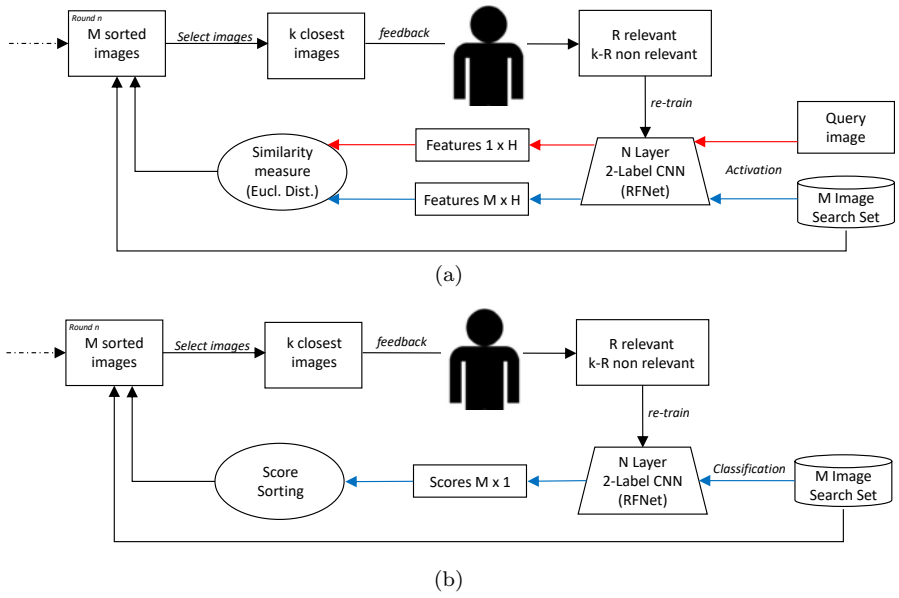

Fig. 4 Conceptual representation of the proposed image retrieval with RF using (a) CNN for feature extraction and (b) the CNN for classification.

In both cases, the new query is used as input to the re-trained CNN instead of the original query. Here we used two different approaches for query shifting. The first one uses the original query image and the images labelled by the user as relevant to compute the most central image, that from now on we will call *KImage* (Eq. (1)). This approach, even if very simple, allows us to identify the image that contains most of the semantic concepts that are relevant to the user. The second one instead directly uses the feature vectors, and it computes a new feature vector as the average of the features extracted from the query image and the relevant images, that from now on we will call *MeanF* (Eq. (2)). Even this approach is very simple, but, by computing the average of all the relevant images features, it allows us to create a new feature vector that contains all the semantic concepts that are relevant to the user.

$$KImage = \underset{\forall I_j \in I_{R+1}}{\operatorname{argmin}} \sum_{i=1}^{R+1} ||I_i - I_j||_2 \tag{1}$$

$$MeanF = \frac{1}{R+1} \sum_{i=1}^{R+1} features(I_i) \tag{2}$$

where I_{R+1} is the set of images composed by the query image and the relevant images.

Since both the previous approaches focus on exploiting the relevant images labelled by the user, we also used the Relevance Score (RS) (Eq. (6)) in combination with the re-trained CNN to facilitate the separation of relevant images from non-relevant ones, that will be referred to in the following as *"strategy_name + RS"*.

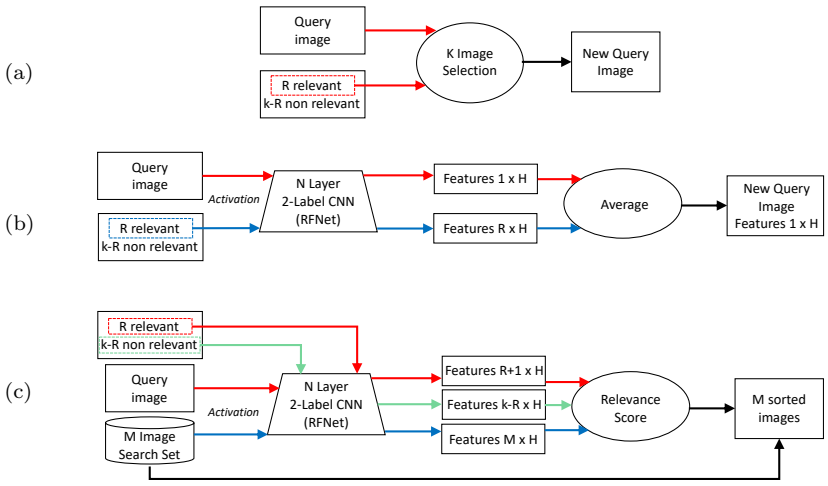

Fig. 5 Representation of the proposed approaches for query refinement: (a) central image extraction, (b) mean features computation and (c) RF using relevance score.

This approach ranks the images in terms of relevance, exploiting both the distance from the relevant images and the non-relevant ones, allowing us to separate these sets of images better. The conceptual representation of the proposed approaches is reported in Figure 5.

5 Materials and Experimental Settings

5.1 Data Sets

We performed several experiments with several data sets differing for the number of classes, and the semantic content of the images.

Caltech is a well-known image data set [1] comprising a collection of pictures of objects. In most images, objects are in the centre with somewhat similar poses, and with very limited or no occlusion. In this work we used the *Caltech-101* and the *Caltech-256* subsets. Caltech-101 is a collection of pictures of objects belonging to 101 categories. It contains a total of 9.144 images, and most of the categories have almost 50 images, but the number of images per categories ranges between 40 and 800. Caltech-256 contains pictures of objects belonging to 256 categories. It contains a total of 30.607 images, and the number of images per category ranges significantly between 80 and 827, with an average value of 100 images per category.

The *Flower* data set[2] presents a collection of flower images. This data set is released in two different versions, and in this work, we used the 102 category version *Flowers-102*. Although this data set has a similar number of classes as Caltech-101, the two data sets are related to two very different problems. Indeed, Flowers-102 turns out to be a problem of fine retrieval, since it contains the single category object 'Flower' that is subdivided into 102 sub-categories. It consists of

[1] http://www.vision.caltech.edu/Image_Datasets/
[2] http://www.robots.ox.ac.uk/ vgg/data/flowers/

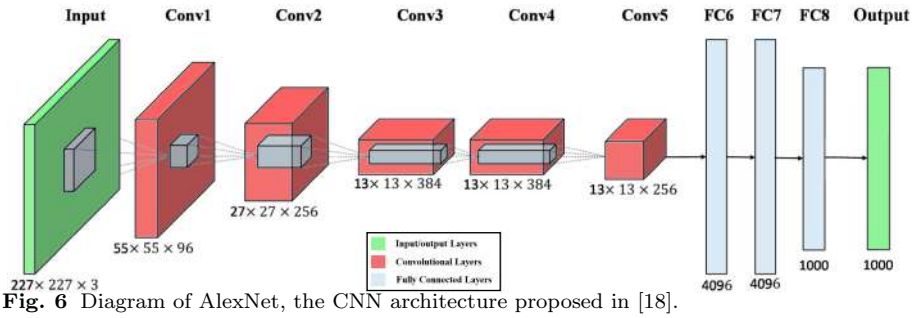

Fig. 6 Diagram of AlexNet, the CNN architecture proposed in [18].

8.189 images, and the number of images per class ranges between 20 and 238. In the experimental evaluation, these three data sets have been divided into two subsets: the query set, containing a query image for each class, and the search set containing all the remaining images for retrieval.

SUN-397 is an image data set for scene categorisation. It contains 108.754 images belonging to 397 categories. The number of images varies across categories, but there are at least 100 images per category. To reduce the data set size, and to balance the number of images in each class, we created the search set, i.e. the set use for retrieval purposes, by taking 100 images per class for a total of 39.700 images. The query set has been created by selecting 250 random samples among the remaining images.

5.2 Pre-trained CNN model

There are several pre-trained models with high popularity thanks to their excellent performances in different tasks. Most of these models are trained with the ImageNet data set[3][11], that is a collection of more than 14 millions of pictures labelled in 21 thousand categories.

AlexNet [18] is one of such models, that is trained on a subset of the ImageNet database, created for the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 [39]. This subset contains 1.2 million training images belonging to 1000 object categories so that on average, each category is represented by almost 1000 images. Although the AlexNet has been trained on a subset of the ImageNet data set, it gained popularity for its excellent performances on many classification tasks. The network architecture consists of 5 convolutional layers, followed by three fully connected layers. In details, the convolutional layers use a different number of kernels, that is 96, 256, 384, 384, 256, respectively, while the size of the kernel gradually decreases from 11x11 to 3x3. Different ReLU, normalisation and pooling layers are inserted within the convolutional layers. The output of the last fully connected layer is sent to a Softmax with loss that produces a distribution over the 1000 image categories (see Figure 6).

[3] http://image-net.org/index

5.3 Setup

The similarity measure used to compute the distances between the query image and all the images in the repository is the Euclidean distance. We evaluated the retrieval performances using the Average Precision (AP) (see Equation (3)) measure since it averages the precision value for all the queries. For a single query i the AP is formulated as follows:

$$AP_i = \frac{1}{Q_i} \sum_{n=1}^{N} \frac{R_i^n}{n} t_n^i \tag{3}$$

where Q_i define the number of relevant images for the i-th query, N is the total number of images of the search set, R_n^i is the number of relevant retrieved images within the n top results, and t_n^i indicates if the n-th retrieved image is relevant ($t_n^i = 1$) for the i-th query or not ($t_n^i = 0$).

We also report the *Precision* (see Equation (4)) that measures the ratio between relevant images within the top k retrieved images.

$$Precision = \frac{Relevant\ Retrieved\ Images}{Retrieved\ Images} \tag{4}$$

To make the RF experiments repeatable and objective we automated the RF process by labelling the retrieval results as being relevant or not according to the matching of each retrieved image with the class label of the query. The underlying assumption is that a user who performs a query should be interested only in images belonging to the same class of the query image. We evaluated the different RF approaches by performing 4 RF rounds for each query image, and by using a number of retrieved images that takes the values $k = 20, 50, 100$. Manually labelling the retrieved images is a very long and time-consuming procedure and no user would like to perform it for more than 20 images, but since the procedure is automated, we also evaluated the higher value of k. This procedure is useful to perform numerical experiments and comparisons between various retrieval systems and to understand if and which system could benefit from a different value of k.

6 Experimental results

We performed several experiments with the previously mentioned data sets, mostly devoted to evaluating the performances of the CBIR system exploiting the user feedback, but firstly we tested the performances of CNN features without the fine-tuning process compared to several state-of-the-art approaches

6.1 CNN features

The CNN features have been extracted from the second fully connected layer (fc7) that produces a feature vector of size $h = 4096$, and they have been compared to the following hand-crafted feature sets. Colour features as proposed in [26] that includes Colour Histogram, Colour Moments and Colour Auto-correlogram, which concatenated produce a feature vector of size 102. SIFT features [23] that have

Table 1 AP on Caltech101, Flowers, Caltech256 and SUN397 by using different feature sets for image retrieval.

Features	Caltech101	Caltech256	Flowers	SUN397
CNN	38.53	18.17	29.81	6.14
Colours	2.96	1.05	5.21	0.48
SIFT	9.99	2.43	4.17	0.83
HOG	8.06	2.33	2.72	0.74
LBP	7.50	2.59	4.11	0.85
LLBP	4.05	1.09	4.24	0.62
HAAR	8.79	2.57	5.56	0.78
Gabor	10.26	2.48	4.27	0.59

been extracted with a grid sampling of 8-pixel size and a window size ranging from 32 to 128. Then the extracted SIFT features have been used to create a BoVW [19] of size 4096. HOG features [8] have been computed on HOG blocks composed by four cells of 16-pixel size. The blocks are overlapped by one cell per side, creating a feature vector of size 6084. LBP [29] have been extracted from blocks of 32-pixel size, to favour the analysis of small regions of the image, since they have been created for texture analysis. The final feature vector has a size of 2891. We also extracted LLBP [37], Gabor Wavelets [41] and HAAR wavelets [44] as are in the original formulation. They present a feature vector of size 768, 5760 and 3456, respectively. In Table 1, we reported the retrieval AP results on each data set. As it can be observed, the retrieval performances obtained are very different for each feature set, but, in general, the use of CNN features outperforms the other approaches in all the tested data sets. These results confirm the trends in image classification already brought to light in [18].

6.2 Results of RF based on CNN

We tested the proposed RF approaches based on CNN on all the data sets previously mentioned. The performances attained with the RFNets have been compared to other RF approaches namely, the Query Shift (QS), the Relevance Score (RS) [15], the Efficient Manifold Ranking (EMR) [50] and a binary Linear SVM classifier [20].

QS is a technique firstly proposed for text retrieval refinement [36] and then adopted in CBIR systems [38]. The assumption behind this approach is that relevant images are clustered in the feature space, but the original query could lie in the region of the feature space that is in some way far from the cluster of relevant images. Accordingly, a new *optimal* query is computed in such a way that it lies near to the Euclidean centre of the relevant images, and far from the non-relevant images, according to Eq.(5)

$$Q_{opt} = \frac{1}{N_R} \sum_{i \in D_R} D_i - \frac{1}{N_T - N_R} \sum_{i \in D_N} D_i \tag{5}$$

where D_R and D_N are the sets of relevant and non-relevant images, respectively, N_R is the number of images in D_R, N_T the number of the total images, and D_i is the representation of an image in the feature space. QS is still widely used in CBIR systems [22,7,21] also by exploiting additional parameters to control

the relative weights of each component. Indeed, in many cases positive feedback is more valuable than negative; thus most information retrieval systems set less weight to negative feedback, or even some system allows only positive feedback, meaning that the weight of negative feedback is set to zero [22]. Conversely, other related studies have shown the great importance of negative feedback for image retrieval [7, 21], and since that the analysis of these parameters is not within the scope of this work, we used the original formulation as in Eq.(5) which places the same importance to both components.

The RS belongs to the Nearest Neighbour (NN) methods used to estimate the posterior probabilities of an image as being relevant or not. NN approaches have been adapted in several forms over the years, but the RS is the most used and still effective form to compute a score for each image [34]. It uses the image distances to its nearest relevant and non-relevant neighbours as follows:

$$rel_{NN}(I) = \frac{||I - NN^{nr}(I)||}{||I - NN^{r}(I)|| + ||I - NN^{nr}(I)||} \tag{6}$$

where $NN^r(\cdot)$ and $NN^{nr}(\cdot)$ denote the nearest relevant and non-relevant image for the image I respectively, and $||\cdot||$ is the metric, typically the Euclidean distance, defined for the feature space.

EMR [50] belongs to Manifold Ranking (MR) approaches, which are graph-based ranking models. Differently from classical MR approaches, that use a traditional k-nearest neighbour graph, the EMR uses k-means for graph construction and a new form of adjacency matrix that optimises the ranking function by least square regression. Although MR approaches are not designed for RF, it turned out that they can handle the feedback very efficiently [50].

The Linear SVM belongs to the Pattern Classification paradigm. The selection of the SVM hyper-parameters has been performed using an error-correcting output code (ECOC) mechanism [2] with 5-fold cross-validation to fit and automatically tune the hyper-parameters. The trained SVM is used like the RFNet to classify the images belonging to the repository, and the resulting score vector can be used as a similarity measure, directly indicating the relevance of an image. The results of this experiment are reported in Figures 7 and 8. As it can be observed, in all the approaches the performances increase after each RF round, but with very different trends, depending on the data set, the size and the number of classes, and the value of k. Indeed, the RFNet approaches heavily depend on the size of the retrieval set k, as it serves as the training set.

It can also be observed that the RF approach based on CNN Features starts converging after few RF rounds, while the RF approaches based on CNN Classification continue improving, even if with different trends depending on the size of k and also with a little lag on the firsts round. This is mainly due to the final layer, introduced during the fine-tuning procedure, that being "new" needs more examples/rounds to adapt to the new task. In contrast, the layer used for feature extraction (fc7) that belongs to the original Alexnet architecture, can immediately provide more representative features. This initial advantage of the RF approach based on CNN Features decreases with the progress of the RF rounds, both because the feature space does not undergo radical changes during the following rounds, and also because the RF procedure depends on the choice of the initial query.

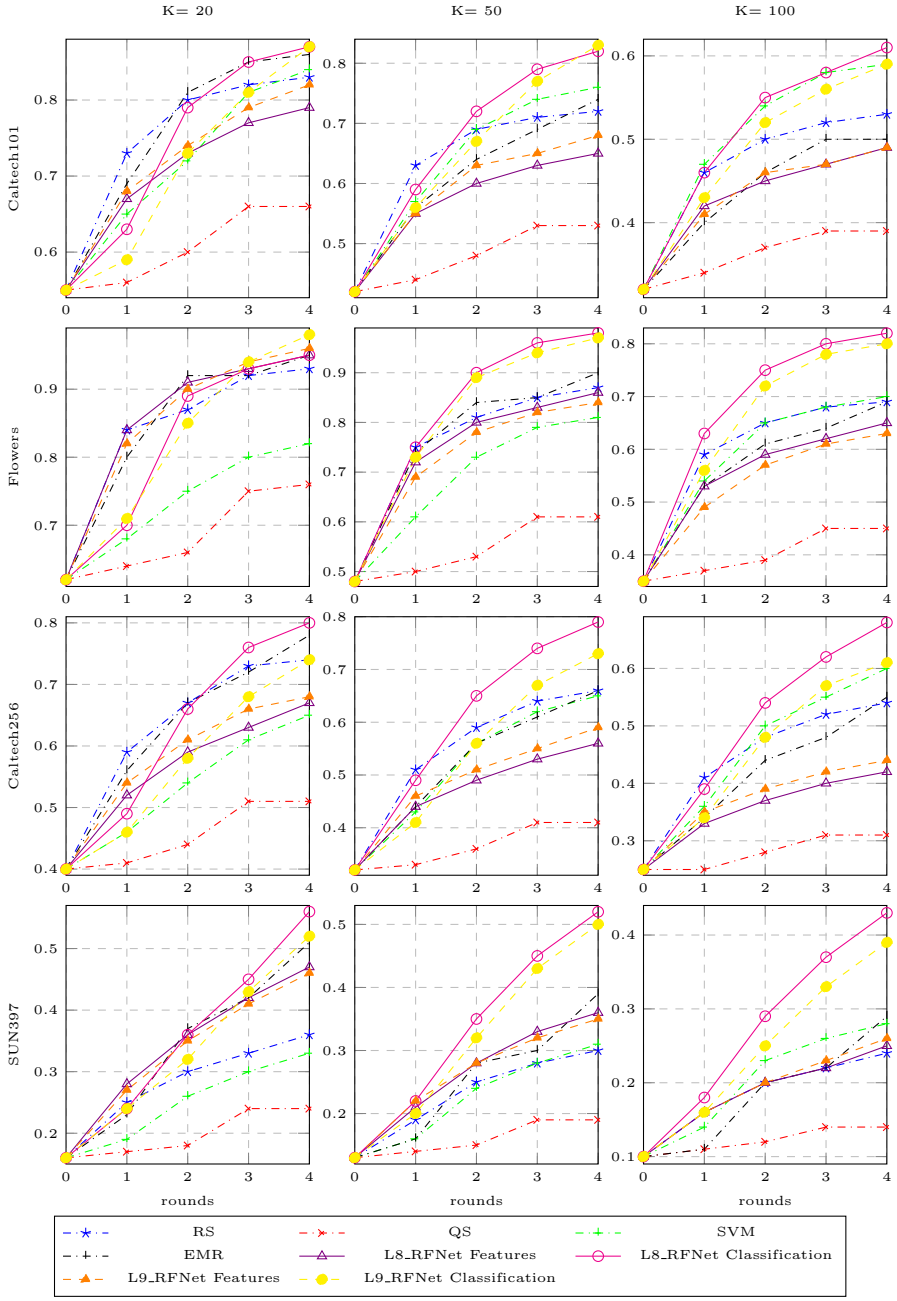

Fig. 7 Precision on Caltech101, Flowers, Caltech256 and SUN397 by exploiting 4 RF rounds for image retrieval on top-k images, with $k = 20, 50, 100$.

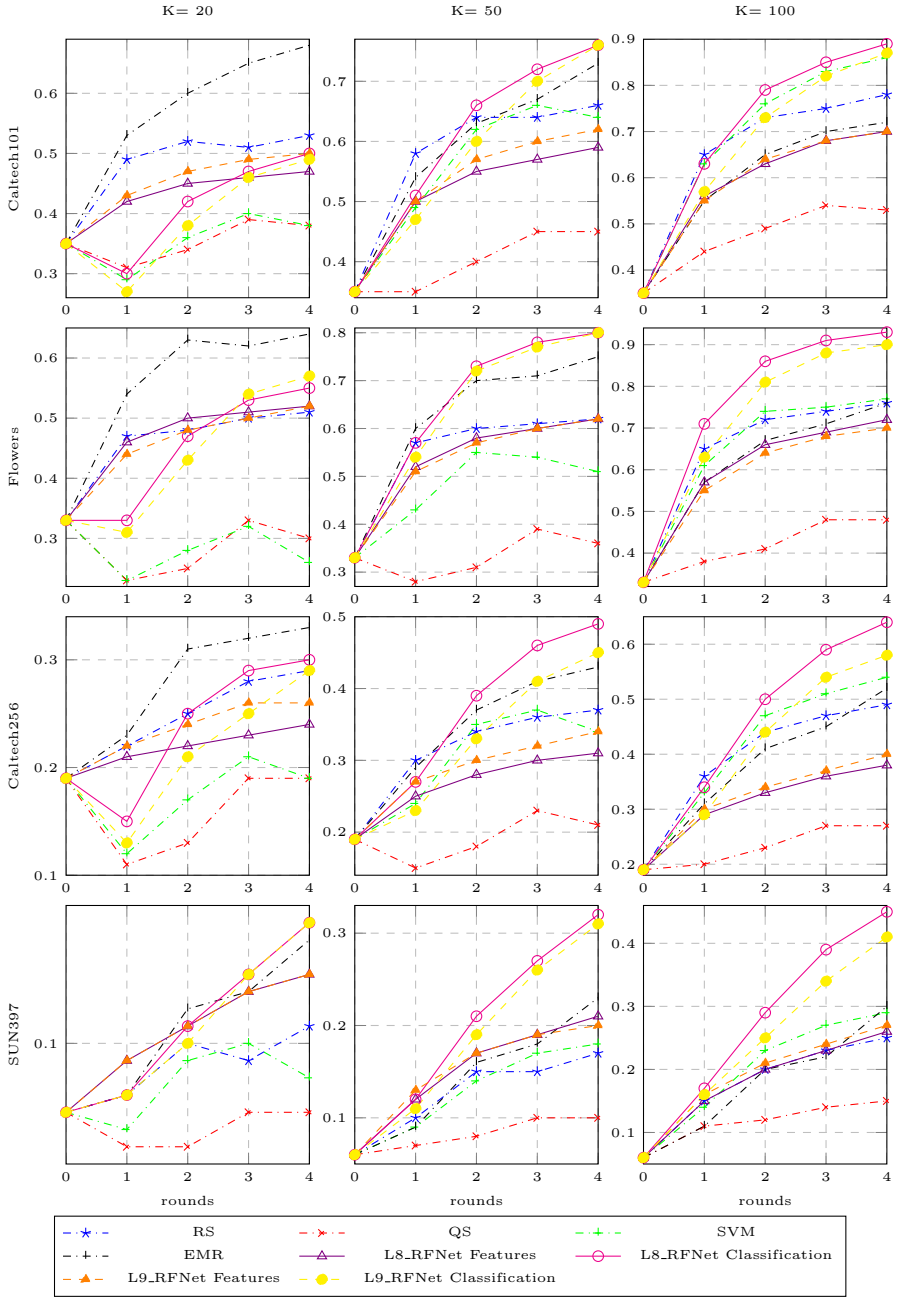

Fig. 8 AP on Caltech101, Flowers, Caltech256 and SUN397 by exploiting 4 RF rounds for image retrieval on top-k images, with $k = 20, 50, 100$.

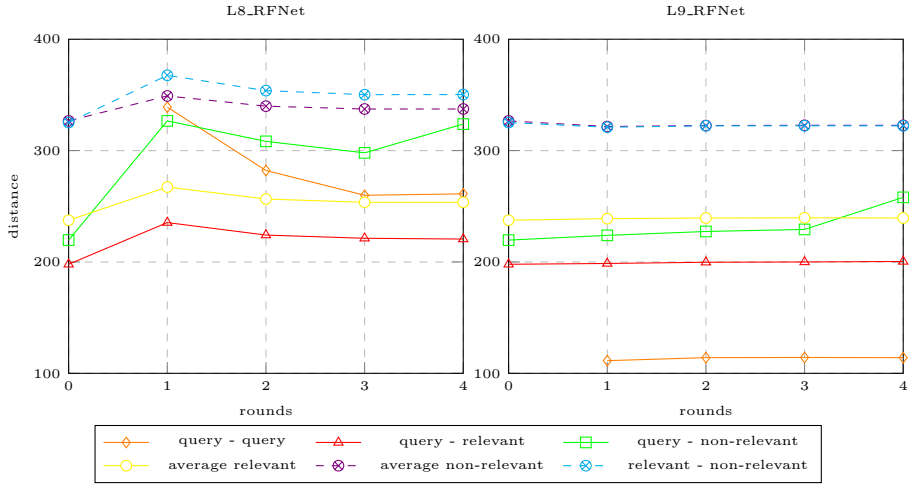

Fig. 9 Feature space evolution: distances between group of features after each re-training respectively on L8_RFNet and L9_RFNet.

7 Results analysis and discussion

To understand the behaviour of the feature space after the re-training of the CNN, we analysed the distances between the query images and the images labelled by the user. Indeed, if the assumption that the relevant images come closer to the query image and far from the non-relevant images is correct, we should see a clear trend after each RF round. Thus, after each RF round and each re-training, we measured the Euclidean distance between the query image features and the relevant and non-relevant images features and also the average distance between relevant and non-relevant images. To show how CNN's internal image representation changes after each re-training phase, we also measured the distance between the original query image features and the new query image features. This results are reported in Figure 9. As it can be observed the image representation changes after each RF round. Nevertheless, there is not a clear trend showing that the two feature sets (relevant and non-relevant) move in different directions, but rather they move in the same direction. This trend is mainly due to the absence of the re-training process before round 0, in which the features were extracted from a generic network, producing a much more generic image representation. Indeed, also the query image features extracted after fine-tuning are quite far from the ones extracted from the original network. These plots are not representative of every single image (monitoring the features of each image could be interesting but not very significant), since they are mediated for all the images labelled by the user on the various data sets, but they clearly show how the network fits new images.

7.1 RF strategy analysis

It is worth noting that the quality of retrieval results is limited by the representativeness of the query image, which often does not reflect the entire set of concepts

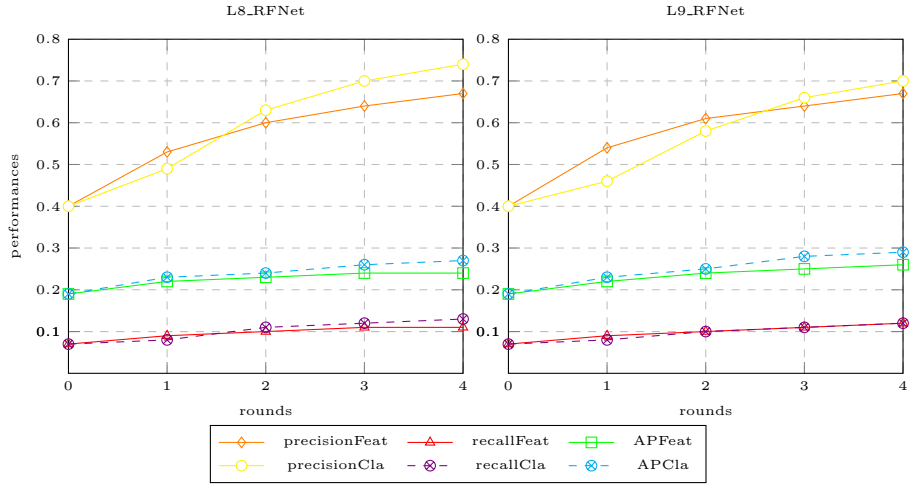

Fig. 10 Comparison between RF approaches exploiting the same net for features extraction and classification.

searched by the user. In the feature space, this translates in the query image being misplaced or marginal with respect to other relevant images. To highlight this behaviour, we compared a single network on both RF strategies, namely the RF based on feature extraction and RF based on classification. The training set has been created using just the images retrieved by the RF feature extraction strategy. Thus the RF classification strategy is slightly disadvantaged by this procedure since it cannot exploit all the images it has retrieved in the previous step. Nevertheless, as it can be observed in Figure 10, it achieves better performances on both the RFNet architectures. Mainly because the classification strategy exploits all the knowledge that the network has learnt during the training process. Instead, using the feature extraction strategy, we exploit just the layer activations, as they are compared to the ones obtained from just one image, i.e., the query.

7.2 Refinement

Thus, although the CNN re-training process aims at modifying the image representation so that relevant images are closer to the query image, the user feedback can be further exploited to refine the query image. To better understand if the RF based on feature extraction could benefit from the proposed refinement methods, we performed a comparison on all the data sets The performances attained with this refinements have been compared with the previous formulation of RF using the RFNet features on both architectures. The results of this experiment are reported in Figures 11. As it can be observed both RFNets exploited with the feature extraction strategy benefit from the proposed refinements. In particular, a significant improvement is observed in each round after the first.

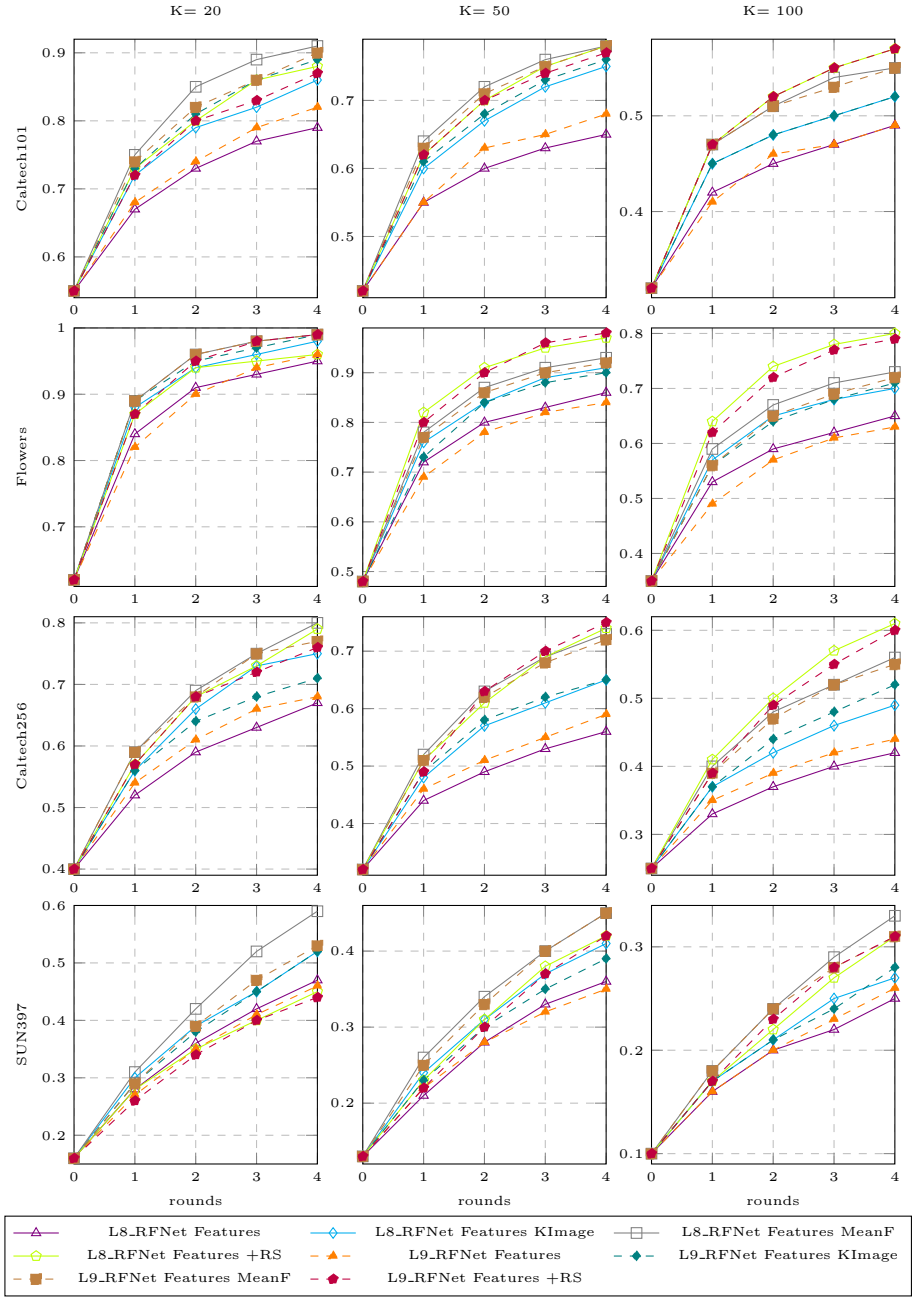

Fig. 11 Precision on Caltech101, Flowers, Caltech256 and SUN397 by exploiting 4 RF rounds with query refinement for image retrieval on top-k images, with $k = 20, 50, 100$.

7.3 Ablation studies

In order to simulate a real case scenario, we performed an experiment where the number of retrieved images at each round is equal to $k = 10$. This is because in many applications, especially those exploiting RF, small windows are used to avoid excessive effort by the user. Moreover, to simulate the variability of human judgement on assessing the relevance of retrieved images to a given query we randomly introduced one or two label flips per retrieval set, by changing the automatic labelling that was used in the previous experiments. The results of these experiments are reported in Figures 12. It is interesting to see that most of the proposed RF approaches achieve high performances even with $k = 10$, thus being effective in exploiting a tiny training set. Indeed, in this case, only the approaches exploiting the CNNs for classification do not perform as well as the others. Nonetheless, all of them performed well even with the presence of label flips, without a performance decrease, but rather all of them provided continuous growth, and even if the performances are not comparable to the baseline (without label flips), this shows remarkable robustness to human variability.

7.4 Overall Results

Table 2 reports the overall results for all the tested RF approaches averaged to all data sets. This table shows the amount of AP gain after each round, and the total amount of gain for each RF approach. We reported in bold the best results obtained on each RF round and in red the results showing a decrease. As a general comment on the attained results, all of the approaches show a performances increase after each RF round, even if with very different trends. In particular, it can be observed that in some RF round the AP decreases when the QS, SVM and RFNet Classification approaches are used. However, while the decrease in AP with the QS and SVM approaches is observed in different RF rounds, the decrease in AP with the RFNet Classification approach is only observed in the first round, outperforming all the other approaches in the remaining RF rounds. Indeed, the RFNet Classification approach does not start converging as quickly as the other approaches, but it is able to improve for other rounds, even if in practice it is not common for users to be engaged in so many RF rounds. In general, the RFNet Classification is the approach that shows the best performances, even if it is heavily dependent on the size k of the retrieval set, as it serves as the training set for the classification layer. Therefore, as expected, the greater the size of the training set (or feedback set) the more we can exploit the learning abilities of CNNs to perform RF. Instead for small sets, it is more effective to use approaches based on features extracted from the RFNet, but only if associated with one of the proposed query refinement strategies such as the Relevance Score and the MeanF. The Relevance Score still exhibits very good performances, but it brings to further improvement when combined with the RF based on RFNet Features on every k value. On the other hand, the MeanF outperformed the other approaches with $k = 10, 20$, which means that the average query feature vector is more meaningful when it is computed from less relevant images.

What emerged could also give rise to hybrid approaches, in fact, given that a single RFNet architecture is used both for feature extraction and classification,

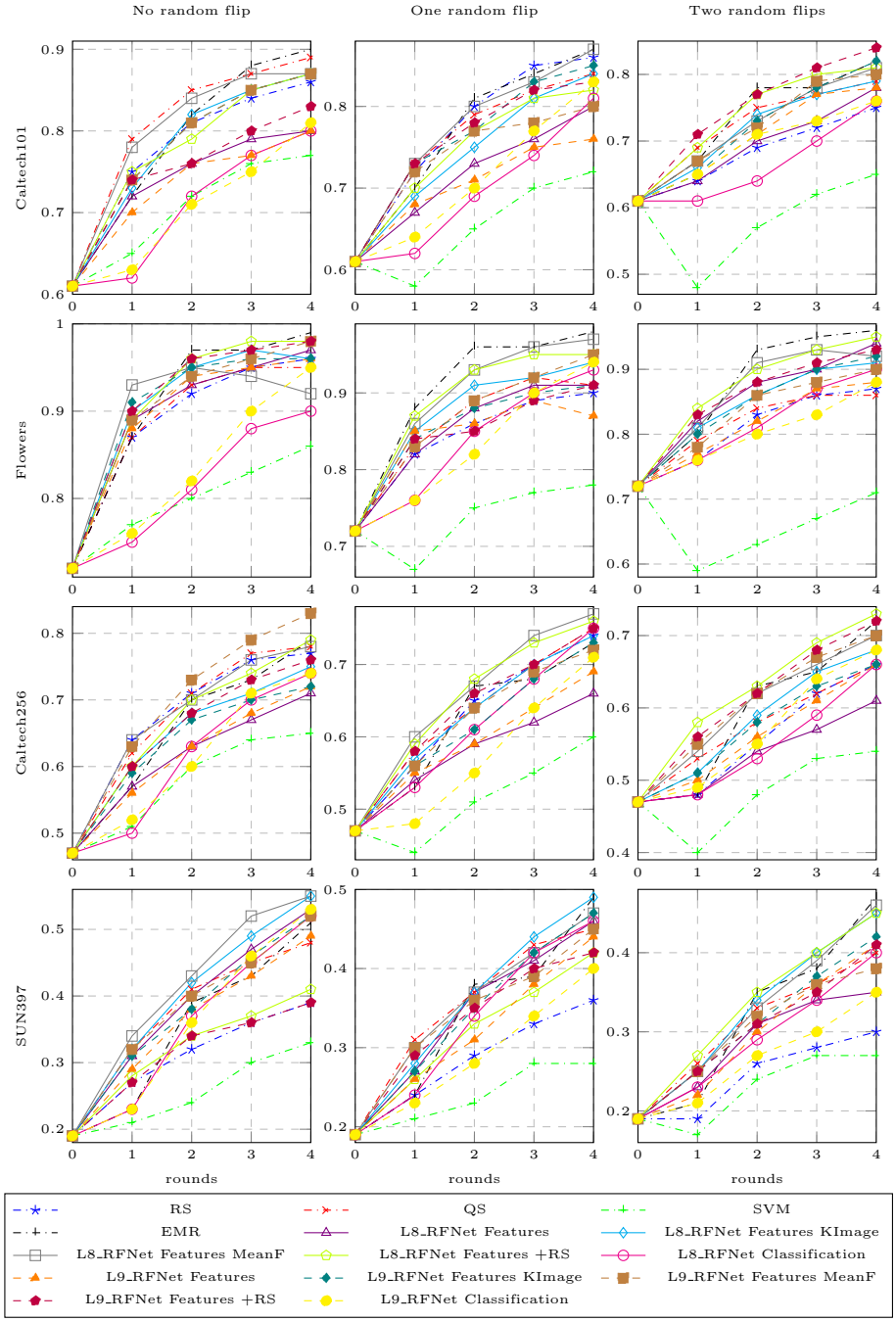

Fig. 12 Precision on Caltech101, Flowers, Caltech256 and SUN397 by exploiting 4 RF rounds for image retrieval on top-10 images, respectively introducing, zero, one or two label flips.

Table 2 Partial and total AP gain, averaged to all data sets, on top-k images, with $k = 10, 20, 50, 100$.

Mode	N. Images	Round				Total Gain
		1	2	3	4	
Relevance Score	10	3.74	2.08	1.17	0.08	7.07
	20	7.73	2.51	1.18	1.07	12.49
	50	15.36	4.38	0.92	1.55	22.21
	100	21.73	7.19	2.37	2.42	33.70
Query Shift	10	7.12	-0.42	-2.95	-3.11	0.64
	20	-5.90	1.50	5.36	-1.41	-0.44
	50	-1.81	2.48	5.24	-1.23	4.67
	100	4.98	2.81	4.50	0.08	12.37
SVM	10	-11.36	2.81	1.59	-3.11	-10.06
	20	-5.89	5.06	3.20	-3.16	-0.79
	50	7.69	10.70	1.66	-1.54	18.51
	100	19.28	12.36	4.05	2.44	38.14
EMR	10	7.63	6.08	0.73	0.64	**15.08**
	20	11.06	7.24	1.40	2.36	**22.05**
	50	14.39	8.89	2.59	4.32	30.19
	100	14.98	9.93	3.77	**5.63**	34.31
L8_RFNet Features	10	4.04	1.14	1.05	0.61	6.84
	20	5.87	2.80	1.15	0.85	10.67
	50	11.66	4.45	2.26	1.66	20.03
	100	16.04	6.08	3.55	2.41	28.08
L8_RFNet Features KImage	10	5.18	2.47	1.58	0.42	9.65
	20	9.42	3.53	1.80	1.19	15.95
	50	15.69	5.81	3.25	2.31	27.05
	100	19.72	7.09	3.92	2.71	33.45
L8_RFNet Features MeanF	10	**8.40**	1.92	1.55	0.54	12.41
	20	**12.56**	4.79	2.14	1.27	**20.76**
	50	**18.39**	6.68	3.53	2.07	30.66
	100	**22.16**	8.67	4.68	3.13	38.65
L8_RFNet Features +RS	10	-4.55	2.10	1.34	1.15	0.04
	20	8.26	4.81	1.89	1.26	16.22
	50	17.48	7.47	4.34	2.61	31.89
	100	**23.51**	9.58	5.65	3.66	42.40
L8_RFNet Classification	10	-11.21	5.91	3.73	2.63	1.06
	20	-2.10	**9.95**	4.42	2.42	14.69
	50	13.41	**13.06**	6.01	3.29	**35.77**
	100	23.07	14.54	7.55	4.51	**49.67**
L9_RFNet Features	10	2.96	1.82	1.01	0.68	6.47
	20	6.10	3.04	2.13	0.99	12.26
	50	11.59	5.08	2.51	1.87	21.05
	100	15.58	6.62	3.59	2.40	28.19
L9_RFNet Features KImage	10	5.36	1.94	1.76	0.96	10.02
	20	9.37	4.03	2.10	1.47	16.97
	50	15.13	6.31	3.37	1.90	26.72
	100	19.48	7.89	3.89	3.05	34.32
L9_RFNet Features MeanF	10	6.76	2.62	2.45	0.96	12.79
	20	11.62	4.86	2.40	1.48	20.37
	50	17.36	7.33	3.51	2.53	30.73
	100	20.93	8.75	4.73	3.24	37.64
L9_RFNet Features +RS	10	4.94	1.62	1.44	0.97	8.97
	20	6.98	5.34	2.18	1.55	16.05
	50	16.34	8.50	4.38	2.58	31.81
	100	22.50	9.97	5.99	3.79	42.24
L9_RFNet Classification	10	-11.55	6.24	3.79	2.75	1.22
	20	-4.11	8.85	**6.52**	**3.41**	14.67
	50	10.44	12.34	**7.48**	**4.33**	34.60
	100	17.98	**14.67**	**8.45**	4.65	45.75

further improvement would be exploiting the network in the first rounds as a feature extractor while in the following rounds as a classifier. The main reason is that the final classification layer being "new" needs more examples/rounds to adapt to the new task, while the layer used for feature extraction (fc7) that belongs to the original Alexnet architecture, can immediately provide characteristic features. On the other hand, the proposed strategy based on classification it is not only intended to use all the knowledge that the network has learnt during the training process, but also to speed-up the indexing process, that in this case is based on the classification scores. Thus, considering that the the time needed to classify or extract the features (from the same network) is almost the same, using the classification strategy we can avoid the similarity measure computation. Several approaches can be employed to reduce the computational cost for both strategies. The simplest approach to avoid re-indexing the whole data set is based on limiting the search to the top N images returned in the first iteration, where N can be chosen based on the data set size M (for example as 10% of M [30]). The hypothesis is that the most important and similar images to the query image are already in the top positions of the ranked lists, especially when using CNN features (see Table 1), and that the images that had already been discarded will not be recovered with RF.

8 Conclusion

Given the great success of CNN in image classification and representation, this paper shows the effectiveness of CNN in image retrieval tasks. Not only we provided a comparison of the effectiveness of features extracted from CNNs compared to hand-crafted features, but also we evaluated different mechanisms for improving the retrieval performances by exploiting Relevance Feedback and a CNN previously trained on a large image data set.

In particular, the user's feedback that labels retrieved images as being relevant or not to the query has been exploited to re-train the CNN to provide a binary output. Since CNN has been modified and trained with samples closest to the user's needs, the extracted features allowed retrieving a large number of relevant images. We proposed different approaches to exploit the RF based on two main strategies, namely by using the CNN as a feature extractor and then computing the image similarity, or by adding a classification layer to label the images as being relevant or not. We also proposed different approaches for reformulating the query to be used as input to the CNN, to combine the modified feature representation with a query more closely related to the user's need. All the proposed approaches showed to be suited to provide significant improvement in the retrieval performances in different experimental settings, i.e., with small or large retrieval sets, as well as in the presence of inaccurate feedback.

It is worth to note how the two proposed strategies are somehow complementary. In fact, by using the CNN as a feature extractor, it allows attaining a significant improvement in the first rounds, and it could be useful for those search engines where the user is not expected to be engaged in a time-consuming feedback loop. On the other hand, the RF based on CNN classification allows further performance improvements when the number of iteration increases, provided that the user labels a relatively large number of images to be used as a training set for

refining the CNN. Accordingly, this setting could be useful in a long term learning task.

References

1. Ashraf, R., Ahmed, M., Ahmad, U., Habib, M., Jabbar, S., Naseer, K.: Mdcbir-mf: multimedia data for content-based image retrieval by using multiple features. Multimedia Tools and Applications (2020)
2. a. Bagheri, M., Montazer, G.A., Escalera, S.: Error correcting output codes for multiclass classification: Application to two image vision problems. In: The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), pp. 508–513 (2012)
3. Baldominos, A., Saez, Y., Isasi, P.: Evolutionary convolutional neural networks: An application to handwriting recognition. Neurocomputing **283**, 38–52 (2018)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
5. Bhowmik, N., González, V.R., Gouet-Brunet, V., Pedrini, H., Bloch, G.: Efficient fusion of multidimensional descriptors for image retrieval. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 5766–5770 (2014)
6. Bulò, S.R., Rabbi, M., Pelillo, M.: Content-based image retrieval with relevance feedback using random walks. Pattern Recognition **44**(9), 2109–2122 (2011)
7. Cruz, M.H.M., Vzquez, M.S.G., Acosta, A.R.: Human vision perceptual color based semantic image retrieval with relevance feedback. In: A.A.S. Awwal, K.M. Iftekharuddin, M.G. Vzquez (eds.) Optics and Photonics for Information Processing XII, vol. 10751, pp. 154 – 162. SPIE (2018)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, pp. 886–893. IEEE Computer Society (2005)
9. Dang-Nguyen, D.T., Piras, L., Giacinto, G., Boato, G., Natale, F.G.B.D.: Multimodal retrieval with diversification and relevance feedback for tourist attraction images. ACM Trans. Multimedia Comput. Commun. Appl. **13**(4), 49:1–49:24 (2017)
10. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys **40**(2), 1–60 (2008)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: A large-scale hierarchical image database. In: CVPR, pp. 248–255. IEEE (2009)
12. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, vol. 32, pp. 647–655. JMLR.org (2014)
13. Erkut, U., Bostancioglu, F., Erten, M., Ozbayoglu, A., Solak, E.: Hsv color histogram based image retrieval with background elimination (2019)
14. Escalante, H.J., Hérnadez, C.A., Sucar, L.E., Montes, M.: Late fusion of heterogeneous methods for multimedia image retrieval. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08, pp. 172–179. ACM, New York, NY, USA (2008)
15. Giacinto, G.: A nearest-neighbor approach to relevance feedback in content based image retrieval. In: CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval, pp. 456–463. ACM, New York, NY, USA (2007)
16. Jiang, M., Zhang, S., Li, H., Metaxas, D.: Computer-aided diagnosis of mammographic masses using scalable image retrieval. IEEE Transactions on Biomedical Engineering **62**(2), 783–792 (2015)
17. Khaldi, B., Aiadi, O., Lamine, K.: Image representation using complete multi-texton histogram. Multimedia Tools and Applications (2020)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., pp. 1106–1114 (2012)

19. Li, Z., Yap, K.H.: An efficient approach for scene categorization based on discriminative codebook learning in bag-of-words framework. Image and Vision Computing **31**(10), 748–755 (2013)
20. Liang, S., Sun, Z.: Sketch retrieval and relevance feedback with biased svm classification. Pattern Recognition Letters **29**(12), 1733 – 1741 (2008)
21. Lin, W.C.: Aggregation of multiple pseudo relevance feedbacks for image search re-ranking. IEEE Access **7**, 147553–147559 (2019). DOI 10.1109/ACCESS.2019.2942142
22. Lin, W.C., Chen, Z.Y., Ke, S.W., Tsai, C.F., Lin, W.Y.: The effect of low-level image features on pseudo relevance feedback. Neurocomputing **166**, 26–37 (2015). DOI 10.1016/J.NEUCOM.2015.04.037
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
24. MacArthur, S.D., Brodley, C.E., Shyu, C.R.: Relevance feedback decision trees in content-based image retrieval. In: 2000 Proceedings Workshop on Content-based Access of Image and Video Libraries, pp. 68–72 (2000)
25. Marques, O.: Visual information retrieval: The state of the art. IT Professional **18**(4), 7–9 (2016)
26. Mitro, J.: Content-based image retrieval tutorial. ArXiv e-prints (2016)
27. Mohan Kumar, P., Balamurugan, B.: Relevance feedback base user convenient semantic query processing using neural network. Advances in Intelligent Systems and Computing **652**, 23–30 (2018)
28. Müller, H., Clough, P.D., Deselaers, T., Caputo, B. (eds.): ImageCLEF, Experimental Evaluation in Visual Information Retrieval. Springer (2010)
29. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
30. Pedronette], D.C.G., Almeida, J., [da S. Torres], R.: A scalable re-ranking method for content-based image retrieval. Information Sciences **265**, 91 – 104 (2014)
31. Pinjarkar, L., Sharma, M., Selot, S.: Deep cnn combined with relevance feedback for trademark image retrieval. Journal of Intelligent Systems (2018). DOI 10.1515/jisys-2018-0083
32. Pinjarkar, L., Sharma, M., Selot, S.: Efficient system for color logo recognition based on self-organizing map and relevance feedback technique. Smart Innovation, Systems and Technologies **77**, 53–62 (2018)
33. Piras, L., Giacinto, G.: Neighborhood-based feature weighting for relevance feedback in content-based retrieval. In: WIAMIS, pp. 238–241. IEEE Computer Society (2009)
34. Putzu, L., Piras, L., Giacinto, G.: Ten years of relevance score for content based image retrieval. In: P. Perner (ed.) Machine Learning and Data Mining in Pattern Recognition, pp. 117–131. Springer International Publishing (2018)
35. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: An astounding baseline for recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '14, pp. 512–519. IEEE Computer Society, Washington, DC, USA (2014)
36. Rocchio, J.J.: Relevance feedback in information retrieval. In: G. Salton (ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall, Englewood, Cliffs, New Jersey (1971)
37. Rosdi Bakhtiar Affendi, C.W.S., Suandi, S.A.: Finger vein recognition using local line binary pattern. Sensors **11**, 1135711371 (2011). DOI http://doi.org/10.3390/s111211357
38. Rui, Y., Huang, T.S., Mehrotra, S.: Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans. on Circuits and Systems for Video Technology **8**(5), 644–655 (1998)
39. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
40. da S. Torres, R., Falcão, A.X., Gonçalves, M.A., Papa, J.P., Zhang, B., Fan, W., Fox, E.A.: A genetic programming framework for content-based image retrieval. Pattern Recognition **42**(2), 283 – 292 (2009)
41. Samantaray, A., Rahulkar, A.: New design of adaptive gabor wavelet filter bank for medical image retrieval. IET Image Processing **14**(4), 679–687 (2020)
42. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations (ICLR2014), vol. abs/1312.6229 (2014)

43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Vision and Pattern Recognition (2014). URL https://arxiv.org/abs/1409.1556
44. Singha, M., Hemachandran, K., Paul, A.: Content-based image retrieval using the combination of the fast wavelet transformation and the colour histogram. IET Image Processing **6**(9), 1221–1226 (2012)
45. Sivic, J., Zisserman, A.: Efficient visual search for objects in videos. Proceedings of the IEEE **96**(4), 548 –566 (2008)
46. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
47. Tzelepi, M., Tefas, A.: Relevance feedback in deep convolutional neural networks for content based image retrieval. In: Proceedings of the 9th Hellenic Conference on Artificial Intelligence, SETN '16, pp. 27:1–27:7. ACM, New York, NY, USA (2016)
48. Tzelepi, M., Tefas, A.: Deep convolutional image retrieval: A general framework. Signal Processing: Image Communication **63**, 30–43 (2018). DOI 10.1016/J.IMAGE.2018.01.007
49. Tzelepi, M., Tefas, A.: Deep convolutional learning for Content Based Image Retrieval. Neurocomputing **275**, 2467–2478 (2018). DOI 10.1016/J.NEUCOM.2017.11.022
50. Xu, B., Bu, J., Chen, C., Wang, C., Cai, D., He, X.: EMR: A scalable graph-based ranking model for content-based image retrieval. IEEE Transactions on Knowledge and Data Engineering **27**(1), 102–114 (2015)
51. Yang, Y., Yang, L., Wu, G., Li, S.: A bag-of-objects retrieval model for web image search. In: N. Babaguchi, K. Aizawa, J.R. Smith, S. Satoh, T. Plagemann, X.S. Hua, R. Yan (eds.) ACM Multimedia, pp. 49–58. ACM (2012)
52. Yu, J., Qin, Z., Wan, T., Zhang, X.: Feature integration analysis of bag-of-features model for image retrieval. Neurocomputing **120**, 355 – 364 (2013)
53. Zhalehpour, S., Arabnejad, E., Wellmon, C., Piper, A., Cheriet, M.: Visual information retrieval from historical document images. Journal of Cultural Heritage **40**, 99–112 (2019)