Incorporation of High Level Information in Images Retrieval

Farzaneh Saadati[1,2], Parvin Razzaghi[2] and Farideh Saadati[2]
[1]*Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran*
[2]*Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran*
f.saadati@iasbs.ac.ir, p.razzaghi@iasbs.ac.ir, saadati@iasbs.ac.ir

Keywords: Image Retrieval, Content-based Image Retrieval, Convolutional Neural Network, Similarity Measure.

Abstract: Content Based Image retrieval (CBIR) is one of the active research areas in computer vision. CBRI searches for similar images from large collections of database images, which belong to the same category of the query image. CBIR is an unsupervised approach that only uses the visual content of an image to retrieve similar images. The main contribution of this paper is to utilize high-level information as well as low-level information to retrieve images. The proposed approach has two steps: (i) a first retrieval set of similar images are obtained using low-level information (ii) for the images of the first retrieval set, high-level information are extracted and then images are reordered. To extract high level knowledge, some candidate objects from each image are obtained. Then each candidate object is described using CNN. In our approach, to define similarity measure, corresponding objects between two images are found and then OMDSL distance metric is applied to compute similarity of corresponded objects. We used MSRC-21 and Caltech256 datasets for evaluating the proposed approach. The obtained results show that our approach outperforms comparable state-of-the-art approaches.

1 INTRODUCTION

The goal of Image retrieval is to retrieve most similar images from a database of images which are relevant to the query image. Up to now, many approaches in this research area are introduced. These approaches are divided into two categories: text-based image retrieval and content-based image retrieval methods. The text-based image retrieval is introduced in 1970s. In these approaches, images are annotated by appropriate keywords, captioning or descriptions to the images. Hence, in the test stage, images which have similar keywords with the query image are retrieved. For example, the famous text search engines are ALPIR and GOOGLE. The main drawback of text-based image retrieval approaches is that image annotation is hard, time consuming and subjective. Furthermore, keywords cannot completely explain the visual content of an image. To overcome these difficulties, content-based image retrieval has been proposed in 1990s. In CBIR, visual content of an image such as color, texture, shape and any related knowledge are utilized to retrieve similar images. Many systems can benefit from accurate content based image retrieval. They include, architecture design (Kekre and Thepade, 2008), image classification (Antani, 2002), medical imaging and geographic info system (Müller, 2004), search engines (Kekre and Thepade, 2009), remote sensing field for indexing biomedical images by contents (Sinha, 2001), weather forecast and criminal investigations. In the following, we provide a brief review of several closely related works. Datta et al.(Datta, 2008) provided a comprehensive survey of image retrieval approaches. Zhao et al. (Zhao) proposed a deep semantic ranking based method for learning hash functions that preserve multilevel semantic similarity between multi label images. In (Wang, 2011), for millions of mobile database images, a new Image retrieval method is introduced which uses vocabulary trees. To construct vocabulary trees, a descriptor contextual weighting (DCW) and a spatial contextual weighting (SCW) of local features are introduced.

The performance of the image retrieval systems is highly dependent on image representation. some features in image representation are color features (Jain and Vailaya, 1996), edge features (Jain and Vailaya, 1996), texture features (Manjunath and Ma, 1996), GIST (Oliva and Torralba, 2001),

CENTRIST (Wu and Rehg, 2011), and the bag-of-words (BoW) models (Wu and Hoi, 2011) using local feature descriptors (e.g. SIFT (Lowe, 1999), and SURF (Bay, 2006),).

In recent years, feature representation using deep learning has received much attention. Researches have shown that features are extracted from the fully-connected layers perform worse than the features that are extracted from the deep convolutional layers of CNNs (Cimpoi, 2015). Convolutional neural networks (CNN's) have been used to learn how to match for the task of stereo estimation (Zagoruyko and Komodakis, 2015). Han et al. (Han, 2015) used a deep convolutional network in a Siamese architecture followed by a fully connected network that learns a comparison function. Zbontar & LeCun (Zbontar and LeCun, 2015) trained CNNs for narrow-baseline stereo and obtained the top results on the KITTI benchmark. These approaches rely on larger networks and do not necessarily learn compact, discriminative representations, compared to ours. In contrast, we show how to exploit discriminative features for image retrieval.

In this paper, a new approach to image retrieval system is proposed which utilizes high-level information as well as low-level information to retrieve most similar images to the query image. The main contribution of this paper are as follows: (i) our approach is done in two level of hierarchy from retrieving a primary coarse similar set of image to a fine similar set of images (coarse to fine retrieval) (ii) incorporating high level knowledge (objects) in the retrieving system (iii) proposing a new similarity measure in the presence of high level knowledge.

The rest of this paper is organized as follow. In section 2, our proposed approach to image retrieval is given in detail. Section 3 shows the results of applying our proposed approach to the best well known MSRC-21 and Caltech 256 datasets. Concluding remarks are given in section 4.

2 PROPOSED APPROACH

In this section, we explain the overall process of our proposed approach. We first apply a pre-retrieval method to the query image and retrieve a large set of images with a global descriptor (e.g. Gist descriptor) which it is explained in detail in section 2.1. Then, the first retrieved large set of images is reconsidered to retrieve most similar images as a final retrieval set. To do this, we extract candidate set of objects with objectness measure (Alexe, 2012) (section 2.2). In the next step, to describe the extracted candidate objects, the pre-trained Convolutional neural network (CNN) (Krizhevsky, 2012) is used (section 2.3). Finally, we use Hungarian method (Kuhn, 1955) to correspond each object in the train image to an object in the query image and then utilize a corresponding matrix to obtain final similarity score and re-rank gained retrieved images set.

2.1 Pre-retrieval

Let $\{I^d\}_{d=1}^{\mathrm{D}}$ be a database images and I^d be a query image. Each image in the database, contains at least one object. In this step, to retrieve a large similar set of images for query image I^q, we apply Gist descriptor (Oliva and Torralba, 2001) to describe each image. The descriptor of query image is denoted by $X^q \epsilon \mathbb{R}^{\mathrm{N}}$ and the descriptor of d^{th} database image (I^d) is denoted by $X^d \epsilon \mathbb{R}^{\mathrm{N}}$ where N is the dimension of the descriptor. To compute the distance measure between query image and dataset image, the Chi square (χ^2) distance (Vedaldi and Zisserman, 2012) is used which is calculated as follows:

$$\chi^2(I_d, I_q) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\left(X_i^q - X_j^d\right)^2}{\left(X_j^d\right)} \qquad (1)$$

Next, we sort images in decreasing order based on their distance measure and top k results are taken. The pre-retrieval step acts like filtering, because it can filter out images that do not have any similarity to the query image. In this step, a large set of images is retrieved. In the following steps, these first retrieved images are reconsidered again to get the final set of retrieval set.

2.2 Extract Candidate Objects

In this step, some candidate objects are taken from each image of the first retrieval set. To get the candidate object from an image, many approaches are introduced. Alex et al. (Alexe, 2012) presented an approach which is called "objectness measure". It refers to a score of how likely a candidate window contains an object of any category. In their work, at first, many candidate windows are sampled randomly and then for each candidate window, a score is calculated based on combination of multiple cues such as saliency, color contrast, edge density, location and size statistics, and how much such windows overlap with super pixel segments.

Manen et al. (Zagoruyko and Komodakis) introduced a fast algorithm for object extraction which is called Prime's algorithm.

In Prim's algorithm superpixels are obtained for each image. For each superpixel, a probability is computed which utilizes from neighboring superpixels belong to the same object. To compute this probability, three cues of color similarity, common border ratio and size are considered. Then proposals are combined using random partial spanning trees. In this paper, to obtain candidate objects, objectness measure (Alexe, 2012) is used.

The extracted candidate objects of the query image I^q are represented by $\{O_i^q\}_{i=1}^{n_q}$ where n_q denotes the number of candidate objects and O_i^q indicates i^{th} bounding box of the query image. Also, the extracted

Candidate objects of database image I^d are represented by $\{O_j^d\}_{j=1}^{n_d}$ where n_d denotes the number of candidate objects and O_j^d indicates j^{th} bounding box of the database image.

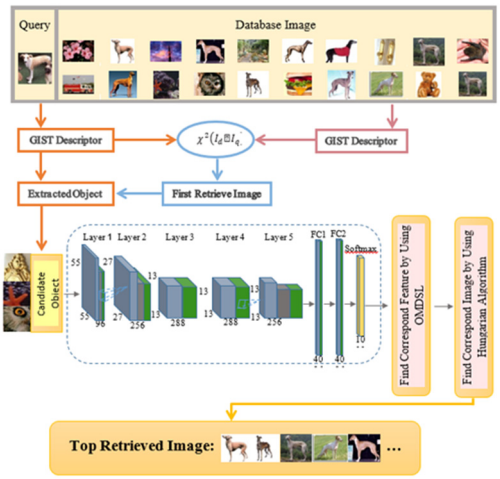

Figure 1: Overall schematic of the proposed approach. In CNN candidate objects are fed to the network as an input layer. Then we have five convolutional layers and three max pooling layers which are represented respectively by blue units and purple units. The green units correspond to the outputs of ReLU transform. Also it has two fully connected layers and an output layer (1000 class labels).

2.3 Feature Extraction

In this section, each candidate-bounding box which is extracted in the previous step, is described by the pre-trained deep convolutional neural networks (CNNs) (Krizhevsky, 2012). In the following subsection, we provide a detail explanation about the used CNN.

2.3.1 Deep Convolutional Neural Networks

In this paper, we use the structure of the CNN in (Krizhevsky, 2012) as pre-trained deep CNN. The structure of Krizhevsky's CNN contains eight learned layers include five convolutional layers, two more fully connected layers (as FC6 and FC7) and the softmax layer. We use the responses from the trained CNN as generic features for each candidate-bounding box.

The first and the second convolution layers of (Krizhevsky, 2012) are followed by a response normalization layer and a max pooling layer. While in the third, fourth, and fifth convolution layers pooling or normalization layer is not used. Krizhevsky's CNN (Krizhevsky, 2012) do better than previous CNNs because the rectified Linear units (ReLUs) is used as a neuron output function, which reduces the training time of the deep CNNs several times more than $f(x)=\tanh(x)$ or $f(x) = (1+e^x)^2$.

The ReLU is applied to the output of every convolution and fully connected layer. Following the convolutional layers, there are two more fully connected layers (as FC6 and FC7) with 4,096 neurons. The last output layer, which is fed by the FC7 layer, is a 1000-way softmax layer, which produces a distribution over the 1,000 class labels in ImageNet. For training of Krizhevsky's deep CNN (Krizhevsky, 2012) the ImageNet's ILSVRC-2012 training set is used, which contains about 1.2 million images.

2.4 Re-ordering

In this step, a new similarity measure is defined for each pair of query image and first retrieval set images. To do this, at first, the corresponding objects between the query image and the database image are found. In this paper, to find corresponding objects, Hungarian algorithm (Kuhn, 1955) is utilized in which to obtain the correspondence matrix between the query image I^q and the database image I^d, the following energy function should be minimized:

$$H(\pi) = \sum_{i,j} C\left(DO_i^q, DO_j^d\right) \qquad (2)$$

Where DO_i^q and DO_j^d denote the descriptor of i^{th} bounding box of the query image I^q and j^{th} bounding box of the database image I^d, respectively. Function $C(.,.)$ represents the distance measure which any distance measure, such as L1, L2, χ^2 and histogram intersection, can be used, however, in our experiment, we adopt the online distance metric

learning algorithm with cosine similarity proposed in (Wu, 2013). In particular, OMDSL method explores a unified two-stage online learning scheme that consists of (i) learning a flexible nonlinear transformation function for each individual modality, and (ii) learning to find the optimal combination of multiple diverse modalities simultaneously in a coherent process. Finally, we sort the images in decreasing order based on our proposed distance measure and then the top *n* results are taken.

3 EXPERIMENTAL RESULTS

In this section, the proposed approach is evaluated. To evaluate our approach, it is applied to MSRC-21 and Caltech-256 datasets. MSRC-21 dataset images are divided into train and test sets based on the standard split method, and they contain 21 categories of classes. Caltech-256 datasets contain 256 categories of classes. In our experiment, 10 classes of Caltech 256 are randomly chosen. Also, it contains 30,607 images. We split the images of each category into 62% as dataset images and 38% as test images. In our approach, to measure the performance of the system, precision is used which denotes the ability of the system in retrieving similar related images. Precision for each query image is calculated as follows:

$$p(i) = \frac{1}{n}\sum_{j=1}^{n} \delta(l(i) == l(j)) \qquad (3)$$

Where n is the number of retrieved images and $l(i)$ and $l(j)$ denote the category label of query image and j^{th} retrieved image. Function $\delta(.)$ maps to 1 at non-negative points, otherwise it maps to 0. In our approach, 10 top-ranked images for performance are considered (n=10). Finally, the average precision is used as performance measure.

As it is mentioned, each candidate object is described by the pre-trained deep convolutional neural networks (CNNs) (Krizhevsky, 2012). We design an experiment to find out which layer can provide a better descriptor for each candidate object. To do this, outputs of different layers are concatenated and used as descriptor, and then retrieval is carried out. Next, for each combination, the average precision is computed (see Table 1). As it is shown in Table 1, FC7 (the second fully connected layer) achieves the best performance with precision of %81 when 10 top ranked images are retrieved. In Figure 2, we show a qualitative evaluation of our approach on two query images from Caltech256 and MSRC-21.

In Table 2, the total average precision of our approach on 10 classes on caltech256 database are shown and are compared with spatial pyramid matching (SPM) method (Lazebnik, 2006), a base line approach (Gist descriptor) and unsupervised bilinear local hashing (UBLH) method (Liu, 2015) in which the first retrieval set of our approach is used as the final retrieval set. The proposed approach has a superior performance compared to the the other methods. It should be noted that our approach, SPM and GIST implemented on 10 random classes of caltech256 and all MSRC-21 database images.

Table 1: Image Retrieval Performance on Caltech256.

Conv1	Conv2	Conv3	Conv4	Conv5	Fc6	Fc7	Softmax	Precision
✓								%30.01
✓	✓							%32.29
✓	✓	✓						%32.29
✓	✓	✓	✓					%32.29
✓	✓	✓	✓	✓				%33
✓	✓	✓	✓	✓	✓			%33
✓	✓	✓	✓	✓	✓	✓		%36.2
✓	✓	✓	✓	✓	✓	✓	✓	%41.08
							✓	%67
						✓		%81
					✓			%75
				✓				%56

Table 2: Image Retrieval Performance on Caltech256 and MSRC-21.

Dataset		Method	SPM (Lazebnik, 2006)	GIST (Oliva and Torralba, 2001)	UBLH (Liu, 2015)	Our approach		
						Fc6	Fc7	Softmax
10 classes	Caltech256	mAP	19%	13%	23.9%	75.47%	81.85%	67.89%
21 classes	MSRC-21		23%	15%	---*	79.33%	83.58%	73.44%

*In the UBLH method also is used SUN397 dataset that mAP is 12.2%.

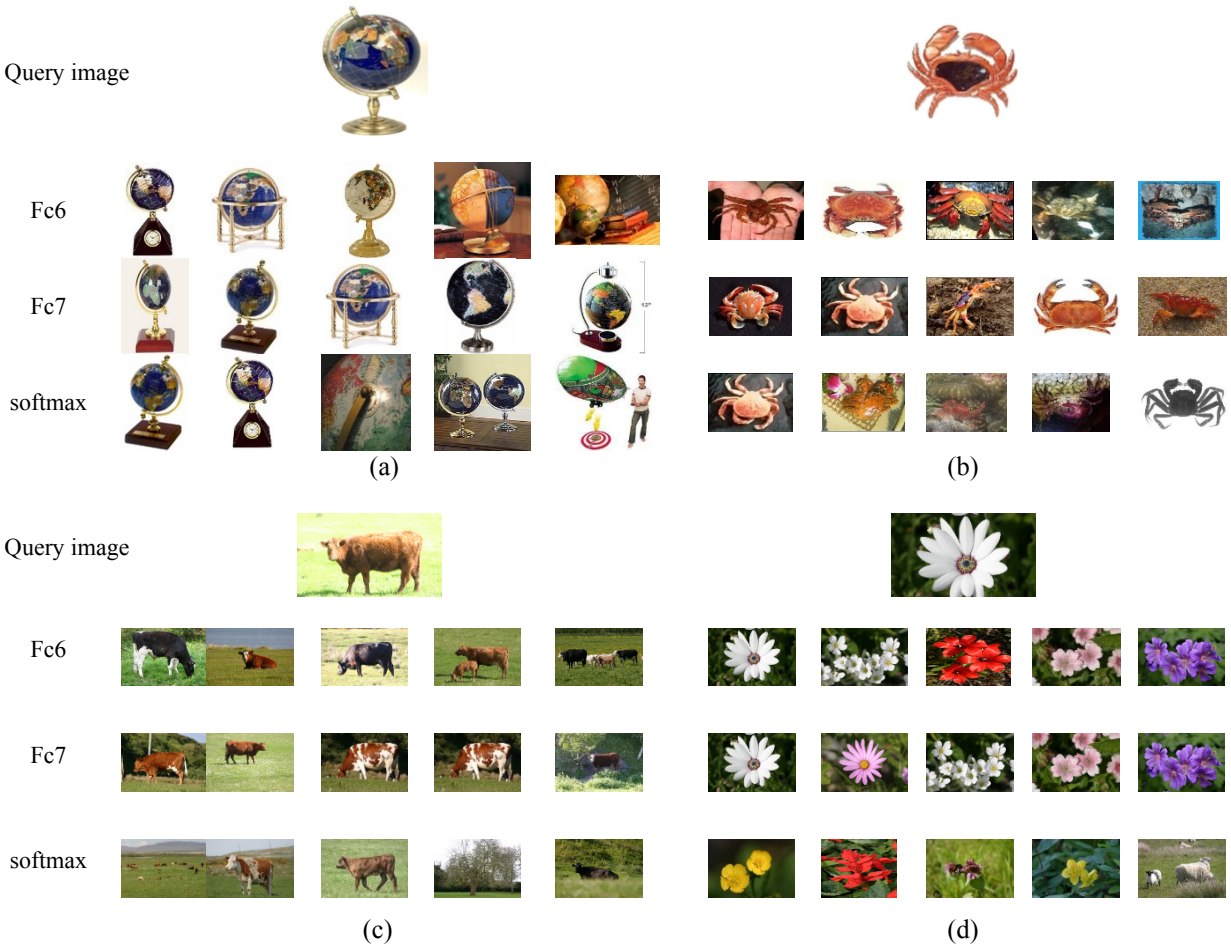

Figure 2: Qualitative evaluation of image retrieval results on Caltech256 ((a) and (b)) and MSRC-21 ((c) and (d)). An image in the above of each row is query, and the images on each row are the top-5 returned results in our method.

4 CONCLUSIONS

This paper proposes an effective method to corporate high-level information in image retrieval. The high-level information denotes the semantic conception of an image like objects. Our suggested method is a coarse to fine retrieval system which in the first step, most coarse similar images are retrieved using Gist descriptor. Then high-level information obtained by extracting objects from each image with objectness measure. In the next step, features are extracted for each object by a convolutional neural network. The extracted features by the CNN model are better than the traditional hand-crafted features. Finally, we used Hungarian algorithm to obtain the correspondence objects between query image and database image. Hungarian algorithm is selected to find object

correspondence, because it is fast and simple. The obtained results show that the proposed method gives better results than GIST algorithm and SPM.

In future works, we will investigate more advanced deep learning techniques and evaluate other more diverse datasets.

REFERENCES

Alexe, B., T. Deselaers ,V. Ferrari, 2012. Measuring the objectness of image windows. Journa.l 34, 2189-2202.

Antani, S., R. Kasturi ,R. Jain, 2002. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Journa.l 35, 945-965.

Bay, H., T. Tuytelaars ,L. Van Gool, 2006. Surf: Speeded up robust features, European conference on computer vision. Springer.

Cimpoi, M., S. Maji ,A. Vedaldi, 2015. Deep filter banks for texture recognition and segmentation, Proceedings of the IEEE Conference on CVPR.

Datta, R., D. Joshi, J. Li ,J. Z. Wang, 2008. Image retrieval: Ideas, influences, and trends of the new age. Journa.l 40, 5.

Han, X., T. Leung, Y. Jia, R. Sukthankar ,A. C. Berg, 2015. MatchNet: Unifying feature and metric learning for patch-based matching, Proceedings of the IEEE Conference on CVPR.

Jain, A. K. ,A. Vailaya, 1996. Image retrieval using color and shape. Journa.l 29, 1233-1244.

Kekre, H. ,S. D. Thepade, 2008. Creating the Color Panoramic View using Medley of Grayscale and Color Partial Images. Journa.l 2.

Kekre, H. ,S. D. Thepade, 2009. Improving the performance of image retrieval using partial coefficients of transformed image. Journa.l 2, 72-79.

Krizhevsky, A., I. Sutskever ,G. E. Hinton, 2012. Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems.

Kuhn, H. W., 1955. The Hungarian method for the assignment problem. Journa.l 2, 83-97.

Lazebnik, S., C. Schmid ,J. Ponce, 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, 2006 IEEE Computer Society Conference on CVPR. IEEE.

Liu, L., M. Yu ,L. Shao, 2015. Unsupervised local feature hashing for image similarity search. Journa.l.

Lowe, D. G., 1999. Object recognition from local scale-invariant features, Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Ieee.

Manjunath, B. S. ,W.-Y. Ma, 1996. Texture features for browsing and retrieval of image data. Journa.l 18, 837-842.

Müller, H., N. Michoux, D. Bandon ,A. Geissbuhler, 2004. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. Journa.l 73, 1-23.

Oliva, A. ,A. Torralba, 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. Journa.l 42, 145-175.

Sinha, U., A. Ton, A. Yaghmai, R. K. Taira ,H. Kangarloo, 2001. Image Content Extraction: Application to MR Images of the Brain 1. Journa.l 21, 535-547.

Vedaldi, A. ,A. Zisserman, 2012. Efficient additive kernels via explicit feature maps. Journa.l 34, 480-492.

Wang, X., M. Yang, T. Cour, S. Zhu, K. Yu ,T. X. Han, 2011. Contextual weighting for vocabulary tree based image retrieval, Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE.

Wu, J. ,J. M. Rehg, 2011. CENTRIST: A visual descriptor for scene categorization. Journa.l 33, 1489-1501.

Wu, L. ,S. C. Hoi, 2011. Enhancing bag-of-words models with semantics-preserving metric learning. Journa.l 18, 24-37.

Wu, P., S. C. Hoi, H. Xia, P. Zhao, D. Wang ,C. Miao, 2013. Online multimodal deep similarity learning with application to image retrieval, Proceedings of the 21st ACM international conference on Multimedia. ACM.

Zagoruyko, S. ,N. Komodakis, 2015. Learning to compare image patches via convolutional neural networks, Proceedings of the IEEE Conference on CVPR.

Zbontar, J. ,Y. LeCun, 2015. Computing the stereo matching cost with a convolutional neural network, Proceedings of the IEEE Conference on CVPR.

Zhao, F., Y. Huang, L. Wang ,T. Tan, 2015. Deep semantic ranking based hashing for multi-label image retrieval, Proceedings of the IEEE Conference on CVPR.