

# Semantic Image Search From Multiple Query Images

Gonzalo Vaca-Castano  
Center for Research in Computer Vision.  
University of Central Florida  
gonzalo@knights.ucf.edu

Mubarak Shah  
Center for Research in Computer Vision.  
University of Central Florida  
Shah@crcv.ucf.edu

## ABSTRACT

This paper presents a novel search paradigm that uses multiple images as input to perform semantic search of images. While earlier focuses on using single or multiple query images to retrieve images with views of the same instance, the proposed paradigm uses each query image to discover common concepts that are implicitly shared by all of the query images and retrieves images considering the found concepts. Our implementation uses high level visual features extracted from a deep convolutional network to retrieve images similar to each query input. These images have associated text previously generated by implicit crowdsourcing. A Bag of Words (BoW) textual representation of each query image is built from the associated text of the retrieved similar images. A learned vector space representation of English words extracted from a corpus of 100 billion words allows computing the conceptual similarity of words. The words that represent the input images are used to find new words that share conceptual similarity across all the input images. These new words are combined with the representations of the input images to obtain a BoW textual representation of the search, which is used to perform image retrieval. The retrieved images are re-ranked to enhance visual similarity with respect to any of the input images. Our experiments show that the concepts found are meaningful and that they retrieve correctly 72.43% of the images from the top 25, along with user ratings performed in the cases of study.

## 1. INTRODUCTION

The goal of any Image Retrieval system is to retrieve images from a large visual corpus that are similar to the input query. To date, most Image Retrieval systems (including commercial search engines like Google<sup>1</sup>) base their search on a single image input query. Recently, multiple images as input queries have been proposed for image retrieval [1, 2]. In these cases, the objective of the multiple image inputs is

<sup>1</sup><http://images.google.com/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [Permissions@acm.org](mailto:Permissions@acm.org).

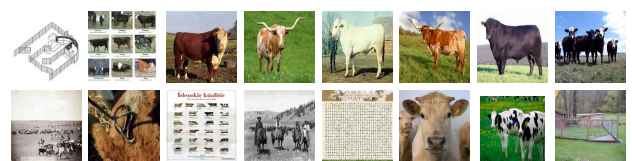
MM'15, October 26–30, 2015, Brisbane, Australia.

© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...\$15.00.

DOI: <http://dx.doi.org/10.1145/2733373.2806356>.



(a) Query Images



(b) Retrieved Images

Figure 1: Results of the proposed paradigm with three input images. a) Query images. b) Top retrieved images.

to acquire different viewing conditions of the unique object or concept for the user.

In this work, we follow a different approach for multiple query image inputs, since the input images are employed jointly to extract underlying concepts common to the input images. The input images do not need to be part of the same concept. In fact, different input image either enriches, amplifies, or reduces the importance of one descriptive aspect of the multiple significances that any image inherently has. Consider the example in Figure 1. Three input images are used to query the system. The first input image is a milk bottle, the second contains a piece of meat, and the third contains a farm. These three images are dissimilar from a visual point of view. However, conceptually, they could be linked by the underlying concept “cattle,” since farmers obtain milk and meat from cattle. These types of knowledge-based conceptual relations are the ones that we propose to capture in this new search paradigm. Figure 1b shows the top retrieved images by our system.

The use of multiple images as input queries in real scenarios is straightforward, especially when mobile and wearable devices are the user interfaces. Consider for instance a wearable device. In this case, pictures can be passively sampled and used as query inputs of a semantic search to discover the context of the user. Once meaningful semantic concepts are found, many specific applications can be built,

including personalization of textual search, reduction of the search space for visual object detection, and personalization of output information such as suggestions, advertising, and recommendations.

The utility of the proposed search paradigm is enhanced when the user has constructed an unclear search or lacks the knowledge to describe it, but does have some idea of images of isolated concepts. For example, consider a user looking for ideas for a gift. While walking through the mall, the user adds some pictures of vague ideas for the gift. Based on the provided images, semantic concepts are found and used to retrieve images of gift suggestions. Another example of this type of application is retrieving image suggestions from a domain-specific knowledge area (painting, cooking, etc.), based on the user's more general visual inputs, where the user does not have specialized knowledge in the specific area. Hence, given a couple of paints that you like as input, the semantic search could find other images for you that share similar common concepts.

This paper presents a new Image Retrieval paradigm with the following novelties:

- 1) Multiple image queries are used as inputs.
- 2) The input images are used to retrieve the concepts that images represent, instead of merely visual similarities.
- 3) Implicit crowdsourcing is used to obtain textual descriptions of pictures that are less noisy than typical surrounding text of web images.
- 4) Text descriptors are used as operands to discover underlying concepts common to the input images. The retrieved images capture semantic similarity from knowledge gained via text concepts and visual similarity.

## 2. METHOD

The proposed method to retrieve images from multiple query input images is explained next. Initially, each one of the query images  $I_i$  is processed individually to retrieve candidate images, based on the visual similarity, from the retrieval dataset. Each query image  $I_i$  produces a set of  $k$  nearest neighbor candidates, denoted as  $C_{ij}$ , where  $j$  goes from  $1 \dots k$ .

All of the images from the dataset used for retrieval are assumed to have a dual representation: a visual representation given by a global image descriptor and a textual descriptor represented by a histogram of all of the texts describing the image. Hence, every candidate image  $C_{ij}$  has an associated textual descriptor represented by a word histogram that serves to link the visual representation with the conceptual representation.

Candidate images that are visually closer to the query image have a higher impact in the text representation of the query image. The most representative text words that describes the query images, are processed using Natural Language Processing (NLP) techniques to discover new words that share conceptual similarity. Afterwards, the histogram summation of the weighted word representations of the candidate images  $C_{ij}$  and the aggregation of bins for the discovered words produces a histogram that represents the queries of the search jointly. Later, cosine distance between the tf-idf textual representation of the database images and the joint search representation is performed to retrieve images. Finally, a re-ranking is performed to privilege images with high visual similarity to any of the query images.

### 2.1 Visual Similarity

We use Convolutional Neural Network (CNN) to generate a high level visual representation of the images. The Convolutional Neural Network (CNN) presented by Krizhevsky et al. [5] contains eight layers with weights; the first five layers are convolutional and the last three layers are fully connected neural networks. The last layer contains 1000 output units fed to a softmax function that determines the class label output. Since we are interested in a high level representation of the image instead of a classifier, we remove the last layer of the network. Features are then calculated by forward propagation of a mean subtracted  $224 \times 224$  RGB image over five convolutional layers and two fully connected layers. The result is a 4096 dimension vector that represents the image as a global descriptor.

Since our image descriptor is global, we perform Locality-Sensitive Hashing (LSH) to have fast approximated nearest neighbor retrieval of candidate images. The presented scheme produces retrievals that are similar in appearance, but also account for the meaning of the image as an object, which is one of the main goals of our proposed framework.

### 2.2 Textual Representation of Dataset Images

The images of the dataset are assumed to have a dual representation (visual / textual). In order to extract textual representation of an image, traditionally the text surrounding the images or associated meta-tags are used as text descriptors for the images. However, surrounding text is highly noisy, and meta-tags are frequently absent, incomplete, or inconsistent with other tasks. We believe that implicit crowdsourcing is a source for cleaner text descriptions. Implicit crowdsourcing engages users in an activity in order to gather information about another topic based on the user's actions or responses. The dataset provided for the Bing challenge [9] is an example of implicit crowdsourcing where text for the images were added by users after they clicked for the best results for their textual searches. Many phrases can describe a single image. Instead of deciding which text best represents the image, we represent the image as a Bag of Words representation of collected text.

### 2.3 Textual Representation of Query Images

The input to our retrieval system is a set of query images without tags, labels, text description, or metadata. In order to operate in the conceptual level, textual representation of the content of the query images must be generated. Textual representations of the top retrieved images from visual features are transferred to obtain a textual representation of each query image. The top retrieved candidates are weighted using a decreasing function of their visual distance to the query image. Hence, if a retrieved image matches exactly with the query image, it receives a maximum weight equal to one, while others receive lower weights. The textual representation of the query image is given by the weighted sum of retrieved textual representations.

### 2.4 Word Representation in Vector Space

The phrases that describe images are morphologically simple. The images in the dataset are fairly well described by a Bag of Words (BoW) textual representation. From a relatively broad vocabulary, only a small set of words are enough to describe the content of images. Hence, similarities be-

```

input :
  • BoW representation of the  $N$  input images
  • Index of the sorted columns of kernel  $K_{ij}$ 
output: List of new words conceptually shared by the
   $N$  input images
initialization;
while  $size(ListNewWords) == 0$  do
  increase number  $T$  of words used to create N-tuples;
  increase number of sorted terms to be intercepted;
  for All  $N$ -tuples from top ranked words do
     $V \leftarrow Intersection(\text{current } N\text{-tuple});$ 
    if  $size(V) \neq 0$  then
      | add  $V$  to ListNewWords
    end
  end
end

```

**Algorithm 1:** Algorithm to discover words conceptually shared by the input images from their textual representations.

tween pairs of words can be enough to capture semantic relations between images.

In the proposed multi-query input image retrieval system, we combine the text descriptor from each input image represented by a small set of words, to infer new words that capture the common meaning of inputs.

Discrete words can be represented in a continuous dense vector space that captures semantic knowledge learned in the text domain. The skip-gram and the Continuous Bag of Words (CBOW) model architectures proposed by Mikolov et al. [7, 8] efficiently learn semantically meaningful float point representations of words from very large text datasets.

Cosine distance can be used in order to find the semantic similarity between two words of the vocabulary, by normalizing the vector representation of each word and computing the dot product between them.

A kernel  $K_{ij}$  that measures the similarity between any word  $i$  and any word  $j$  of the vocabulary is calculated during the training phase. The purpose of the kernel  $K_{ij}$  is to quickly find the similarity of any pair of words.

## 2.5 Concept Discovery

The representation of a query image is typically formed by only tens of words. If  $N$  is the number of query inputs and one word is selected from each query image, an  $N$ -tuple of words is formed. We want to examine the  $N$  words of the  $N$ -tuple to discover new words that relate them conceptually.

A word  $W_i$  is considered as a new word concept, when  $W_i$  has simultaneously high similarity to all the words of the  $N$ -tuple measured in the vector space.

The kernel  $K_{ij}$  enables finding the similarity between any pair of words of the vocabulary. A row (or column)  $q$  of the kernel matrix  $K_{ij}$  is the distance of the word indexed by  $q$  to any other word of the vocabulary. Performing a descend sort operation and taking their indices in the selected column  $q$ , gives the word indices that are conceptually closer to the word  $W_q$ .

Given an  $N$ -tuple, we can find the closest words for all the words that are part of the  $N$ -tuple. If a word  $W_i$  is

found in the top positions of all the sorted lists of the  $N$ -tuple words, then the word  $W_i$  is declared as a new word conceptually shared by the  $N$ -tuple. We call this procedure “intersection.”

There are many combinations of  $N$ -tuples that could be formed; however, the most interesting ones are the  $N$ -tuples created from words that have highest weights in the textual representation of each query image.

A small number  $T$  of words with higher weights are used to create the  $N$ -tuples. When no shared word is found in the given set of  $N$ -tuples, the number of words used to create  $N$ -tuples and the number of sorted terms to be intercepted is increased until at least one common word is found. Algorithm 1 summarizes this procedure.

## 2.6 Image Retrieval

Image retrieval is performed from a unique textual representation that accounts for all input images and the list of words conceptually shared by the input images. The textual representation of the joint query is the summation of the individual query input representations and the addition of new bins indexed by the list of words shared by the input images.

Image retrieval is performed based on the ranking score produced by the cosine scoring algorithm [6] between the joint search representation and the tf-idf representation of the images of the dataset.

Instead of comparing with the entire set of images of the dataset, we use a shorter number of possible outputs to calculate the score of the images. This subset of images is formed by all the images of the dataset that contain at least one word from either the list of words conceptually shared by the inputs or the most representative words of each query input. Hence, the number of images to evaluate reduces from one million to just a couple of thousand candidate images in the Bing dataset. The retrieved images are based on their conceptual ranking only. Therefore, a re-scoring of the retrieved images is performed based on the visual similarity to the input images. Images that are visually inconsistent with all the input images are penalized, and re-ranked to lower positions. The value of penalization is calculated using an inverse exponential function of the Euclidean norm of the difference between descriptors of the retrieved image and its most visual similar input image.

## 3. EXPERIMENTS

We use the public domain implementation and trained model available from the caffe library [4] to calculate our visual image representation. Image resizing and feature extraction of one million images can be performed in less than 24 hours in a regular Quad core personal computer.

The dataset [9] provided for the 2013 MSR-Bing Image Retrieval Challenge<sup>2</sup> was sampled from one-year click logs of the Microsoft Bing image search engine. We chose this dataset for image retrieval since the texts associated with the images are fairly accurate because they are built from user’s search criteria and click preferences. The most important advantage is that the image labeling does not require humans dedicated to this activity and that is the product of

<sup>2</sup><http://acmmm13.org/submissions/call-for-multimedia-grand-challenge-solutions/msr-bing-image-retrieval-scientific-track/>

Table 1: Mean accuracy of the retrieved images according to user ratings in 101 pairs of query images. Results are showed at different top retrieval levels.

Method	Top 5	Top 10	Top 15	Top 20	Top 25
Baseline (Before visual re-ranking)	0.5058	0.4843	0.4823	0.4774	0.4784
Baseline (After visual re-ranking)	0.5549	0.5294	0.5124	0.5093	0.5011
Ours (Before visual re-ranking)	0.7509	0.7450	0.7327	0.7172	0.7098
Ours (After visual re-ranking)	<b>0.7765</b>	<b>0.7579</b>	<b>0.7490</b>	<b>0.7358</b>	<b>0.7243</b>

implicit crowdsourcing. The average number of queries per image is 23.1, consequently there is a significant amount of text describing each image.

A stop list is created to ignore frequent words in the dataset. Words that appear in over 50000 images are ignored since they correspond to non-discriminative words.

The resulting text representation of the images is very sparse since only a few tens of words describe each image. For this reason, text representation can be saved in a very compact way. In fact, the full dataset representation of one million images can be fully loaded in 51 Mb of memory.

For our experiments, we used vector representation of words trained on a part of the Google News dataset, which contains about 100 billion words.<sup>3</sup> The vocabulary size  $D$  of this model is 300.

### 3.1 Multiple Query Image Retrieval

We performed experiments to evaluate our system for a set of 101 pairs ( $N = 2$ ) of image inputs. The definition of the input pairs of images was performed with the help of semantic maps downloaded from the internet.<sup>4</sup>

A semantic map is a visual strategy for vocabulary expansion and extension of knowledge by displaying words and their relations with other words [3]. Any person can define a semantic map about any topic; therefore, there is no “correct” semantic map. We chose several sets of pairs of words that were related according to the semantic maps found. Words that were not easily represented pictorially were discarded, and the remaining words were used to download example images and form query pairs of images conceptually related.

For each pair of available query images, we asked several users to rate the top 25 retrieved images of a pair of query images. They were then asked to provide a binary answer to the following question: Is the retrieved image conceptually similar to both input images or not?

Based on the user ratings, we calculated the mean accuracy of the retrieved images from the 101 pairs of query images. Accuracy is reported for the top  $X$  retrieved images, with  $X$  ranging from 5 to 25 in intervals of 5.

The baseline method is defined as the image retrieval performed from a joint search representation given by the summation of the individual input textual representations without the addition of words conceptually shared by the input images.

Table 1 presents the results of accuracy of our method and the baseline. For both methods we include the results before applying visual re-ranking. Our method clearly outperforms the baseline results by more than 22% in the proposed task. The visual re-ranking also helps to improve the ratings of the users. The improvement is more significant in the base-

line case, where the retrieved images are worse in the task of retrieving shared concepts. The previous observation is an indication that some users tend to rate positively near identical images when the conceptual meaning of the query images cannot be clearly established.

## 4. CONCLUSIONS

We have introduced a new search paradigm that leverages representations of multiple input images to infer concepts shared by all the input images. The retrieved images are conceptually and visually meaningful.

We presented a complete solution to the problem. Our solution includes novel visual and text representation of the images, exploiting a dataset with annotations obtained by implicit crowdsourcing, and operating a vector space that allows conceptually measuring the similarity between words.

The proposed approach achieved mean accuracy of 77.65% on the top 5 retrievals and 72.43% on the top 25, according to user ratings. The proposed approach outperforms the baseline by more than 22%, which shows that the method works very well in the proposed problem.

## 5. REFERENCES

- [1] R. Arandjelovic and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012.
- [2] F. Basura and T. Tuytelaars. Mining multiple queries for image retrieval: On-the-fly learning of an object-specific mid-level representation. In *ICCV*, 2013.
- [3] G. G. Duffy. *Explaining Reading, Second Edition: A Resource for Teaching Concepts, Skills, and Strategies*. Guilford Press, 2009.
- [4] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. In <http://caffe.berkeleyvision.org/>, 2013.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [6] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press New York, 2008.
- [7] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*, 2013.
- [8] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, 2013.
- [9] MSR-Bing. Image retrieval challenge dataset. In <http://web-gram.research.microsoft.com/GrandChallenge/Datasets.aspx>.

<sup>3</sup><https://code.google.com/p/word2vec/>

<sup>4</sup><http://www.bing.com/images/search?q=semantic+mapping>