A Probabilistic Approach for Image Retrieval Using Descriptive Textual Queries

Yashaswi Verma
CVIT, IIIT Hyderabad, India
yashaswi.verma@research.iiit.ac.in

C. V. Jawahar
CVIT, IIIT Hyderabad, India
jawahar@iiit.ac.in

ABSTRACT

We address the problem of image retrieval using textual queries. In particular, we focus on descriptive queries that can be either in the form of simple captions (e.g., "a brown cat sleeping on a sofa"), or even long descriptions with multiple sentences. We present a probabilistic approach that seamlessly integrates visual and textual information for the task. It relies on linguistically and syntactically motivated mid-level textual patterns (or *phrases*) that are automatically extracted from available descriptions. At the time of retrieval, the given query is decomposed into such phrases, and images are ranked based on their joint relevance with these phrases. Experiments on two popular datasets (UIUC Pascal Sentence and IAPR-TC12 benchmark) demonstrate that our approach effectively retrieves semantically meaningful images, and outperforms baseline methods.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.10 [**Computing Methodologies**]: Artificial Intelligence-Vision and Scene Understanding

General Terms

Algorithms, Experimentation, Measurement

Keywords

Image Retrieval, Descriptive Queries, Statistical Models

1. INTRODUCTION

Since the past decade, there has been an outburst of multimedia content, particularly in the form of digital photographs and videos. This has led to new challenges in accurate and efficient archiving as well as retrieval of this content. While most of this content is in free-form, a considerable portion of it is loosely linked with textual meta-data. This has made it feasible to study the associations between the two modalities, which in turn can be useful in semantic analysis of the unannotated visual data [14, 16, 18].

In this work, we address the problem of image retrieval using textual queries. In the conventional text-based image retrieval set-up, a query is usually a set of one or more labels. However, these labels carry minimal linguistic information (such as relationships among labels in terms of action and relative position). As a result, the retrieved images may not match the desired semantics. Hence it comes as a natural choice to develop approaches that can support image retrieval using descriptive queries (e.g., [14]). Such queries inherently carry information about the properties of individual objects as well as relationships among different objects, which can be useful in semantically coherent retrieval.

With this motivation, we present a probabilistic approach for image retrieval using queries that can be either in the form of simple captions, or even long descriptions. Our approach is motivated by the fact that available annotated images (i.e., images with corresponding descriptions) can be harvested to perform retrieval on unannotated images. Assuming a collection of annotated images, our goal is to rank unannotated (test) images based on their semantic similarity with a given query. This similarity seamlessly integrates both visual as well as linguistic aspects. The visual semantics of a test image are approximated based on its relevance with a subset of its most similar images retrieved from the annotated collection. The semantics of the query are determined based on its relevance with the descriptions of the retrieved subset of annotated images as above. One distinctive aspect of our approach is that rather than using either individual words from a query, or the full query as is, images are ranked based on their joint relevance with the *phrases* automatically extracted from the query. These phrases can be thought of as *mid-level textual patterns* corresponding to different visual aspects that may be depicted in an image (e.g., "person near car", "blue airplane", etc.). Experiments validate the intrinsic capacity of our approach in efficiently capturing the semantics of a query, and also demonstrate its superior performance compared to baseline techniques.

2. RELATED WORK

The problem of image retrieval is a well-studied topic in computer vision and multimedia [2]. Two popular streams in this field are: (a) content-based retrieval of images (e.g., retrieve images similar to a given query image [1]), and (b) directly retrieving images based on textual category [7]. In the first setting, the role of semantics of the query image is minimal, and retrieval is usually performed using a standard

MM'15, October 26–30, 2015, Brisbane, Australia
 ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806289.

set of visual features (e.g., GIST). In the second setting, semantics get introduced in the form (textual) labels. Due to this, now it is possible to have two visually dissimilar images depicting the same concept.

The task of image retrieval gets semantically richer as more and more language aspects get introduced into the query. E.g., performing image retrieval for queries that are bigger than a single category label (say a pair of labels) may be more meaningful than a single-label query. This now requires to look for images containing all the concepts rather than just one or few of them. Image annotation methods such as [8, 4], that aim at associating a *set* of relevant labels with a given image, can be useful in multi-label image retrieval. However, one limitation of this setting is that it does not capture the relationships among the labels. E.g., the query "person, car" gives no clue about the relative position of "person" and "car". This gives rise to several possibilities, such as a "person" can be sitting inside a "car", standing near a "car", etc. This limitation was addressed in [15], where it was proposed that it is semantically more meaningful to learn complete phrases in the visual domain (e.g., "person sitting in car") rather than individual labels.

A complementary task to image retrieval is to associate an image with semantically meaningful text (say a group of labels, phrases, or captions). In order to get a deeper understanding of image semantics, it is desirable to associate images with language constructs that are even bigger than phrases (such as captions). With this motivation, several approaches have been proposed that are aimed at describing the content of an image, such as [6, 14, 16, 18]. These descriptions are usually in the form of simple captions containing few tens of words. Among these, there are two popular practices: either to *generate* a description given an image [6, 16, 17], or to *retrieve* one from a collection of available descriptions [11, 14, 18]. In the first setting, a new description is generated by combining visual clues using natural language generation (NLG) techniques. Whereas in the second setting, a description is retrieved by using either image-to-image similarity [11] or image-to-text similarity [14, 18]. The approaches that perform image-to-text similarity are cross-modal retrieval approaches, and have also been shown to be useful in retrieving images given a descriptive query. Lately, there have also been some attempts like [9] that address this task using deep learning models.

The motivation of our work is similar to cross-modal image retrieval approaches [14, 18]. However, one limitation of these is that they represent a complete description (query) using a single feature vector. Due to this, it becomes difficult to understand the impact of linguistic aspects of a query. As we will discuss in the next section, our approach in fact makes use of the linguistic structure of a query, and image retrieval is performed based on these properties. This in turn can be helpful in analyzing the interplay among the individual components of the query. Conceptually, our work closely relates with the image description generation methods [6, 17], and demonstrates their application to the image retrieval task given descriptive textual queries.

3. APPROACH

Our approach consists of two phases. In the first (training) phase, we extract a set of textual phrases from the descriptions of training images, and learn the parameters of our model. In the second (retrieval or testing) phase, we extract all the phrases from a given query, and rank the test images based on their joint relevance with these phrases.

3.1 Training Phase

3.1.1 Phrase Extraction

One of the most important components of our approach is to effectively harness the semantic information encoded in the available descriptions. Rather than considering objects, attributes, verb, preposition, etc. from a sentence in a piece-wise manner, we extract relation tuples or phrases [1]. These carry a bigger chunk of information compared to individual components of a sentence. To extract phrases, the available descriptions are parsed using the Stanford CoreNLP toolkit[2]. As suggested in [3], we use "collapsed-ccprocessed-dependencies" that are useful for extracting relations. From the parsed sentence, a set of phrases of the form $(subject, verb)$, $(verb, prep, object)$, etc. are extracted ("*prep*" stands for preposition). In practice, we extract nine types of phrases: $(subject)$, $(object)$, $(attribute, subject)$, $(attribute, object)$, $(subject, verb)$, $(object, verb)$, $(subject, prep, object)$, $(object, prep, object)$, and $(verb, prep, object)$. We consider two set-ups while extracting phrases: without synonyms and with synonyms. In the first set-up, we consider all the extracted phrases as is. However, in the second set-up, we consider synonyms by expanding each noun (object/subject) up to at most three hyponym levels using its WordNet synsets. This results into replacing a noun with its most popular synonym (e.g., replacing "pug" with "dog", "Ferrari" with "car", etc.).

3.1.2 Phrase Relevance Prediction Model

Let $\mathcal{T}_r$ be the collection of training images that are annotated with descriptions. As discussed above, we extract phrases from all the available descriptions, and then map each description to a set of phrases. Let $\mathcal{Y}$ denote the collection of all the extracted phrases. Then the training data takes the form $\mathcal{T}_r = \{(I_i, Y_i)\}$, where $Y_i \subseteq \mathcal{Y}$ is the set of phrases corresponding to image I_i. Each image is represented using a set of n features $\{f_1, \ldots, f_n\}$ such as colour histograms, texture features, etc. For parameter learning, the training set $\mathcal{T}_r$ is divided into disjoint training and validation sets $\mathcal{T}$ and $\mathcal{V}$ respectively. Now, given an image $I \in \mathcal{V}$, we compute the relevance of a phrase $y_i \in \mathcal{Y}$ with it using their joint probability score $P(y_i, I)$ defined as ([8]):

$$P(y_i, I) = \sum_{J \in \mathcal{T}_I^K} P_{\mathcal{T}}(J) P_{\mathcal{F}}(I|J) P_{\mathcal{Y}}(y_i|J) \quad (1)$$

Here, $\mathcal{T}_I^K \subset \mathcal{T}$ is the set of K nearest neighbours of I from $\mathcal{T}$. These neighbours are determined based on distance of I from the images in $\mathcal{T}$ in the feature space. Let J be an image from $\mathcal{T}$, then distance between I and J is defined as:

$$D_{I,J} = w_1 d^1_{I,J} + \ldots + w_n d^n_{I,J} = \mathbf{w} \cdot \mathbf{d}_{I,J} \quad (2)$$

Here $d^1_{I,J}, \ldots, d^n_{I,J}$ are distances between the corresponding features $f_1, \ldots, f_n$ of both the images computed using some specific distance metric (e.g., Manhattan or Euclidean distance), $w_1, \ldots, w_n$ are non-negative real-valued weights denoting linear distance combination, and $D_{I,J}$ is the actual distance between the two images I and J.

[1]For simplicity, we shall refer "relation tuples" as "phrases".
[2]http://nlp.stanford.edu/software/corenlp.shtml

As shown in Eq. 1, there are three components in our definition of joint probability. The first component $P_{\mathcal{T}}(J)$ denotes the probability of picking an image J from $\mathcal{T}_I^K$. Assuming that all the neighbouring images are equally likely, this is modeled as a uniform prior; i.e., $P_{\mathcal{T}}(J) = \frac{1}{K}$. The second component $P_{\mathcal{F}}$ denotes the likelihood of *seeing* image I given its neighbouring image J, and is defined as a function of distance between the two images:

$$P_{\mathcal{F}}(I|J) = \frac{\exp(-D_{I,J})}{\sum_{J' \in \mathcal{T}_I^K} \exp(-D_{I,J'})} \quad (3)$$

The above definition signifies that the weights will decay smoothly with distance, which in turn helps in adjusting the distance during parameter learning.

The last component in Eq. 1 denotes the probability of seeing the phrase y_i given image J, and is defined as [4]:

$$P_{\mathcal{Y}}(y_i|J) = \frac{\mu\delta_{y_i,J} + N_i}{\mu + N} = \frac{\nu\delta_{y_i,J} + (N_i/N)}{\nu + 1} \quad (4)$$

Here, $\nu = \frac{\mu}{N} \geq 0$ is a smoothing weight, N_i denotes Google count[3] of the phrase y_i, and N denotes the total Google count of all the phrases. Since our approach is data-driven, and the number of phrases is usually large, it is difficult to predict their general behaviour using only the available data. Hence, Google counts help in estimating the statistical behaviour of different phrases.

Let Y_J^t be the phrases in Y_J that are of the same *form* as that of y_i (e.g., say both are $(subject, verb)$). Then, $\delta_{y_i,J}$ in the above equation is given by:

$$\delta_{y_i,J} = U_{sim}(y_i, Y_J^t) = \max_{y_j \in Y_J^t} Z_{sim}(y_i, y_j) \quad (5)$$

where $Z_{sim}(y_i, y_j) \in [0, 1]$ denotes similarity between the phrases y_i and y_j. This is computed using the procedure described in [17], which is based on WordNet based similarity between the individual terms of the two phrases. The above definition denotes that if the phrase y_i is present in the ground-truth phrases of J, then $\delta_{y_i,J} = 1$. Otherwise, it will be the similarity score of y_i with the closest matching phrase in Y_J^t. Such a definition helps in considering semantic interdependence among phrases while predicting phrase relevance rather than just presence/absence.

From Eq. 1, we can observe that there are two types of parameters in our phrase relevance model: distance weights w_i and smoothing weight ν. Given an image $I \in \mathcal{V}$, and $Y_I \subset \mathcal{Y}$ being the phrases extracted from its descriptions, our goal is to learn the above parameters such that (a) the probability of predicting any phrase $y_j \notin Y_I$ should be small, and (b) the probability of predicting a phrase $y_i \in Y_I$ should be more than that of predicting any other phrase $y_j \notin Y_I$. With this goal, our loss function is defined as:

$$e = \sum_{I,y_j} P(y_j, I) + \lambda \sum_{(I,y_i,y_j) \in \mathcal{M}} \eta_{ij}(P(y_j, I) - P(y_i, I)) \quad (6)$$

where $\lambda > 0$ takes care of the trade-off between the two competing terms, $\eta_{ij} = 1 - Z_{sim}(y_i, y_j)$, and $\mathcal{M}$ is the set of triples (I, y_i, y_j) (with $y_i \in Y_I$ and $y_j \notin Y_I$) that violate the second constraint as discussed above. The significance of η_{ij} is that if two phrases are semantically similar (e.g., "bus on road" and "coach on highway"), then the penalty for giving higher score to the phrase not present in the ground-truth (y_j) should be small, and vice-versa. To optimize the parameters, we use a stochastic gradient descent method.

[3]The number of approximate search results obtained using Google search for an exact match query.

3.2 Retrieval Phase

In the previous section, we described the model for predicting the relevance of a phrase with a given image. This is done by combining the similarity of the given image with available images, and similarity of the given phrase with those in the ground-truth of available images. Here we describe how this model can be used for performing retrieval on an unannotated image collection given a descriptive query.

Let $\mathcal{T}_e$ be the collection of (unannotated) test images. These images constitute our retrieval set. To perform retrieval on this set, we make use of the available images with descriptions. Note that now we consider all the samples in $\mathcal{T}_r$, which was earlier partitioned into training and validation sets for parameter optimization.

During the retrieval phase, we are given a descriptive query Q, and our goal is to rank the images in $\mathcal{T}_e$ based on their relevance with the query. This is done based on the posterior $P(J|Q)$ of an image $J \in \mathcal{T}_e$ given the query Q:

$$P(J|Q) = \frac{P(Q, J)}{P(Q)} \propto P(Q, J) \quad (7)$$

To compute $P(Q, J)$, first we parse the query and extract all its phrases (Y_Q) as described in Section 3.1.1. Then for each phrase $y \in Y_Q$, we compute its relevance score with the given test image J. This score is the joint probability of associating y with J, and is obtained using Eq. 1. Here we consider the annotated images from $\mathcal{T}_r$ to compute the neighbouring images of J, and pick the K most similar images. These images are then used to compute the different components ($P_{\mathcal{F}}(\cdot)$ and $P_{\mathcal{Y}}(\cdot)$) of Eq. 1. After computing the relevance of all the phrases in Y_Q with J, we compute the joint relevance score of associating J and query Q as:

$$P(Q, J) = \prod_{y \in Y_Q} P(y, J) \quad (8)$$

Similarly, we compute the joint relevance scores for all the images in $\mathcal{T}_e$, and then rank them in the descending order of this score (higher score means more relevance).

It is worth noticing that the second term in Eq. 1 ($P_{\mathcal{F}}(\cdot)$) considers only visual similarity, i.e., query-independent. Hence, this needs to be computed just once for all the images in the retrieval set, and can be done off-line. In other words, we can pre-compute the K nearest neighbours $\mathcal{T}_J^K \subset \mathcal{T}_r$ of each test image J and store their indices and conditional probability scores. Then, during the retrieval phase, we need to compute just the relevance of the phrases in the given query with those of the images in $\mathcal{T}_J^K$ (using Eq. 4).

4. EXPERIMENTS

We evaluate and compare our approach on two popular image description datasets: (1) **UIUC Pascal Sentence:** This was introduced in [13], and is a de facto benchmark for evaluating image-caption associations [16, 6, 18]. It contains 1,000 images, each annotated with captions from 5 human-annotators. On an average, each description has around 10 words. (2) **IAPR-TC12 benchmark:** This was introduced in [5] for cross-language retrieval, and has 19,627 images. Each image is annotated with a long description of up

Dataset →	Pascal		IAPR-TC12	
Method ↓	BLEU-1	Rouge-1	BLEU-1	Rouge-1
CCA [14]	0.29	0.17	0.26	0.28
BITR [18]	0.31	0.21	0.27	0.26
Ours (syn.✗)	0.33	0.22	0.31	0.30
Ours (syn.✓)	0.36	0.24	0.35	0.33

Table 1: Comparison of our approach to the state-of-the-art methods using automatic evaluation.

Figure 1: Sample queries from the Pascal Sentence dataset along with the top two retrieved images.

to 5 sentences. Compared to Pascal Sentence dataset, this has more diverse object categories and complicated descriptions, thereby providing a challenging test-bed for evaluation. On an average, each description has around 25 words.

We represent each image using a set of global and local features similar to [6]. The global features include GIST, colour histograms in RGB and HSV colour spaces, and Gabor and Haar features. The local features include bag-of-words histogram using SIFT descriptor. Except GIST, we also compute other features over three vertical and horizontal partitions of an image and concatenate them. This helps in encoding the spatial layout of an image into the features. While computing distance between two images, Euclidean distance is used for GIST, Gabor and Haar features, Manhattan distance for colour histograms, and chi-square distance for bag-of-words histogram (of SIFT features).

4.1 Experimental Set-up and Comparisons

For evaluations, we partition each dataset into 45% training set, 45% retrieval set and 10% query set. The training set is used to learn the model parameters, the images in the retrieval set constitute the images on which we perform retrieval, and the descriptions in the query set are used for querying the retrieval set. This is repeated ten times in order to include all the descriptions in a dataset into the query set. While extracting phrases from available descriptions, we consider two set-ups, where (1) all the phrases are considered as such, and (2) each subject/object in a phrase is replaced by its synonym determined using WordNet synsets.

For evaluation, we consider BLEU [12] and Rouge [10] metrics. Following [18], for a given query, we average these scores over the top five retrieved images, by matching the query with their ground-truth descriptions. For both these measures, we report average unigram scores.

Since our work is closely related to cross-modal image retrieval methods, we compare with two such methods [14, 18]. Both these methods have been shown to perform well for image retrieval using descriptive queries, and cross-modal retrieval in general. To evaluate both these methods, we follow the same experimental set-up as described in [18].

4.2 Results and Discussion

In Table 1, we report the results for automatic evaluation. Evaluations using both the metrics confirm the superior performance of our approach compared to [14, 18]. This validates that during retrieval, it is better to consider meaningful chunks of a descriptive query (phrases in our case) rather than the whole query as is. The results also show that by considering synonyms, we are able to achieve better performance than without synonyms. This is because nouns play a central role in image retrieval. On replacing each noun with its synonym, the number of distinct phrases reduces. This results in lowering the competition among semantically similar phrases, and thus improves the chances of retrieving images that better match a query. Figure 1 shows the top two images retrieved for sample queries from the Pascal Sentence dataset. Here it can be observed that usually the retrieved images are quite relevant to a given query. E.g., in the third column, none of the retrieved images has a "blue *and* white airplane". However, there is either a blue or a white coloured airplane in each image, with a background depicting the other colour (white building or blue sky). This indicates that even when the image content does not completely match with the query, we are able to retrieve images that match its components.

Acknowledgement: Yashaswi Verma is partially supported by Microsoft Research India PhD fellowship 2013.

References

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.
[2] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences and trends of new age. *ACM Computing Surveys*, 2008.
[3] M.-C. de Marneffe and C. D. Manning. The stanford typed dependencies representation. In *COLING Workshop*, 2008.
[4] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.
[5] M. Grubinger. *Analysis and Evaluation of Visual Information Systems Performance.* PhD thesis, 2007.
[6] A. Gupta, Y. Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.
[7] V. Jain and M. Varma. Learning to re-rank: Query-dependent image re-ranking using click data. In *WWW*, 2010.
[8] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance model. In *SIGIR*, 2003.
[9] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014.
[10] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACLHLT*, 2003.
[11] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2Text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.
[12] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 2002.
[13] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collective image annotation using amazon's mechanical turk. In *NAACLHLT Workshop*, 2010.
[14] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.
[15] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.
[16] Y. Ushiku, T. Harada, and Y. Kuniyoshi. Automatic sentence generation from images. In *ACM MM*, 2011.
[17] Y. Verma, A. Gupta, P. Mannem, and C. V. Jawahar. Generating image descriptions using semantic similarities in the output space. In *CVPR Workshop*, 2013.
[18] Y. Verma and C. V. Jawahar. Im2Text and Text2Im: Associating images and texts for cross-modal retrieval. In *BMVC*, 2014.