Attribute-Graph: A Graph based approach to Image Ranking

Nikita Prabhu and R. Venkatesh Babu
Video Analytics Lab, SERC
Indian Institute of Science, Bangalore, India.
nikita@ssl.serc.iisc.in, venky@serc.iisc.in

Abstract

We propose a novel image representation, termed Attribute-Graph, to rank images by their semantic similarity to a given query image. An Attribute-Graph is an undirected fully connected graph, incorporating both local and global image characteristics. The graph nodes characterise objects as well as the overall scene context using mid-level semantic attributes, while the edges capture the object topology. We demonstrate the effectiveness of Attribute-Graphs by applying them to the problem of image ranking. We benchmark the performance of our algorithm on the ‘rPascal’ and ‘rImageNet’ datasets, which we have created in order to evaluate the ranking performance on complex queries containing multiple objects. Our experimental evaluation shows that modelling images as Attribute-Graphs results in improved ranking performance over existing techniques.

1. Introduction

In a digital world of Flickr, Picasa and Google Image Search, ranking retrieved images based on their semantic similarity to a query has become a vital problem. Most content based image retrieval algorithms treat images as a set of low level features or try to define them in terms of the associated text [15, 38]. Such a representation fails to capture the semantics of the image. This, more often than not, results in retrieved images which are semantically dissimilar to the query.

Image processing and computer vision researchers to date, have used several different representations for images. They vary from low level features such as SIFT [21], HOG [3], GIST [23] etc. to high level concepts such as objects and people [20]. Since we want our image retrieval system to rank images in a way which is compatible with visual similarity as perceived by humans, it is intuitive to work in a human understandable feature space. When asked to describe an object or a scene, people usually resort to mid-level features such as size, appearance, feel, use, behaviour etc. Such descriptions are commonly referred to as the attributes of the object or scene. These human understandable, machine detectable attributes have recently become a popular feature category for image representation for various vision tasks [31, 43, 45]. In addition to image and object characteristics, object interactions and background/context information form an important part of an image description. It is therefore, essential, to develop an image representation which can effectively describe various image components and their interactions.

Graphs, which have long been used by the vision community to represent structured groups of objects, are an ideal tool for this purpose [9, 12, 22, 34]. We represent images as Attribute-Graphs, using graph nodes to represent objects in the image and the global context of the scene. Each object is described using object-specific local attributes, and the overall scene with global attributes, thereby capturing both local and global descriptions of the image. Apart from the objects of importance in the scene, our model also incorporates the geometrical layout of the objects in the image. The edges of the graph capture the location, structure and orientation of the nodes with respect to other nodes. The proposed Attribute-Graphs thus characterise an image from three different perspectives, namely,

Figure 1. Same objects, yet different semantics: (a) A dog next to a person lying on a sofa (b) A person and a dog near a forest (c) A dog watching a person from a distance

Figure 2. (a) An example query image (b) & (c) Expected Retrieved images: Ones with same semantics
global (scenes), local (objects) and sub-local (attributes) as well as the inter-relations between the same. This allows them to better conceptualise the essence of the image.

We then use graph matching to ascertain the similarity of each of the dataset image models to the query model. The pipeline of the proposed method is shown in Fig. 3. The goal of our work is to rank images according to their semantic similarity to the query, by developing an image representation capable of unifying the varied aspects in an image.

The rest of the paper is organised as follows. In Section 2 we detail the prior work in this area. In Section 3, we discuss crucial aspects of the concept of semantic similarity of images. Keeping the aforementioned aspects in mind, we describe the construction of Attribute-Graphs in Section 4. In Section 5, we describe our datasets ‘rPascal’ and ‘rImageNet’ created to evaluate image ranking. The performance of the proposed algorithm is discussed in Section 6. Section 7 concludes the paper with a summary of our work.

2. Related Work

Mid level features have long been used for various tasks in the field of computer vision. The recent surge in the use of attributes have reiterated the efficacy of such features. Farhadi et al. [8] showed the usefulness of attributes for image description, object classification and abnormality detection. Extending the applicability of attributes beyond objects, Patterson et al. [25] demonstrated their use for scene description and classification. Attributes have also been employed for image understanding [26, 39], web search [13, 14], zero-shot-learning [11, 18], action recognition [32] and human-computer interaction [24].

In addition to these applications, attributes have recently been increasingly used for the task of image retrieval and ranking. As attributes can effectively describe mid-level semantic concepts, they are a convenient means to express the users’ search intentions. However most attribute based image retrieval works [35, 41] focus on textual queries. For example, Siddiquie et al. [35] exploit the interdependence among both query and non-query attributes, and model the correlation among them to improve retrieval performance. Yu et al. [41] improve upon the performance of [35] by modelling query attribute dependency on a pool of weak attributes. Images however, are an amalgamation of several constituents, and cannot be accurately described by just a few words. It is therefore difficult to capture the complete gist of an image merely by textual descriptions [35, 41], structured queries [19], concept maps [40] or sketches [7]. Such techniques tend to focus on one or two aspects of the image while ignoring the rest. While Lan et al. [19] consider the spatial relationship between a pair of objects, they do not account for the overall geometrical layout of all the objects and the object characteristics. Xu et al. [40] while maintaining spatial relationships do not consider background information and object attributes. Kulkarni et al. [17] consider both objects and their interrelations, but do not model the background holistically.

Content based image retrieval techniques such as [6, 42] use image queries. Zheng et al. [42] couple complementary features of SIFT and colour into a multidimensional inverted index to improve precision, while adopting multiple assignment to improve recall. Douze et al. [6] use attributes in combination with Fisher vectors of a query image to perform retrieval. These techniques obtain a single global representation for an image, and fail to consider the objects in the image and their local characteristics. Cao et al. [1] perform image ranking by constructing triangular object structures with attribute features. However, they fail to take into account other important aspects such as the global scene context. We compare the proposed method with the works of Douze et al. and Cao et al. in Sec. 6.

3. Ranking Algorithm: Aspects to Consider

People tend to look for objects in images. Therefore, visually similar images would generally contain the same objects. Consider the images in Fig. 1. All the three images, contain the same objects, namely a person and a dog. Yet they are semantically very different from each other. This illustrates that mere presence of similar objects is insufficient to make two images similar.
4. Attribute-Graphs for Image Ranking

4.1. Object detection and classification

Objects play an important role in our understanding of an image. Therefore, accurate detection and classification of the objects is essential to construct a good representation for an image. Convolutional Neural Network (CNN) based algorithms have recently shown an improved performance over most other techniques for the tasks of detection and classification of objects [4, 10, 16]. We employ the algorithm of Girshick et al. [10] to localise and classify objects. Girshick et al. obtain region proposals for an image by performing selective search as described by Uijlings et al. [37]. A high-capacity CNN is then utilised to obtain 4096 dimensional features for the obtained region proposals. Class specific linear SVMs are used to score these proposals. Given all scored regions in an image, a greedy non-maximal suppression is used to eliminate low scoring proposals and obtain the final object regions $R_1, R_2, \ldots$ etc., and their corresponding classes $o_1, o_2, \ldots$ etc.

4.2. Attribute-Graph construction

The proposed method uses undirected fully connected graphs $G(V,E)$ to represent images. Here, $V = \{V_l, V_g\}$ represent the nodes and $E$ represents the set of edges connecting the nodes. Each object present in the image contributes to a graph node resulting in a total of $N$ object nodes or local nodes denoted by $V_l = \{v_1, \ldots, v_N\}$. The additional node $V_g$ (also referred to as the global node) represents the background or the overall scene. An image with $N$ objects is thus transformed into a graph having $N + 1$ nodes. The edge sets of the graph are defined as follows:

- Local edges: Edges between two object nodes. There will be $\binom{N}{2}$ such edges.
- Global edges: Edges between the object nodes and the global node ($V_g$). $N$ such edges exist.

The following sections (4.2.1 & 4.2.2) describe the features used to characterise the Attribute-Graph.

4.2.1 Node features

Each object node is represented using object level attributes (local attributes). These local attributes are limited to the area occupied by the bounding box of that particular object and represent the semantic characteristics of that object. The global node captures the overall essence of the image. We use a different set of attributes, which we refer to as global attributes to define this node. These attributes are extracted from the entire image ($I$) and describe the image as a whole. This enables the global node to represent the image context or scene characteristics effectively. The node features are assigned according to Eq. (1).

$$
\phi(v_j) = \begin{cases} 
\varphi(I) & \text{if } v_j = V_g \\
\psi(v_j) & \text{otherwise (1)}
\end{cases}
$$

$v_j$ represents the $j^{th}$ node, $\psi$ extracts the local attributes of the object and $\varphi$ is a function which returns the global attributes of the image. In Fig. 4, the object nodes dogs and sofa are described by local attributes such as has head, is wooden etc., and the scene by global attributes such as living room, interior etc.

4.2.2 Edge features

The edge features of the model are defined so as to capture the spatial configuration of the image components. The local edges capture the relative arrangement of the objects with respect to each other while the global edges define the positioning of the objects in the image. The edge features are represented by Eq. (2).

$$
\chi(e_{ij}) = \begin{cases} 
[\mu_{ij}, \theta_{ij}, o_{ij}] & \text{if } v_i, v_j \in V_l \\
[\mu_{ig}, \theta_{ig}, area(v_i)] & \text{if } v_i \in V_l \& v_j = V_g (2)
\end{cases}
$$

$e_{ij}$ represents the edge connecting node $v_i$ to node $v_j$. $\mu_{ij}$ is the pixel distance between object centroids. $\theta_{ij}$ represents the angle of the graph edge with respect to the horizontal taken in the anti-clockwise direction, while ensuring
left/right symmetry. It indicates the relative spatial organisation of the two objects. Left/right symmetry is ensured by considering an angle $\theta$ to be equal to the angle $(180 - \theta)$. $\alpha_{ij}$ represents the amount of overlap between the bounding boxes of the two objects and is given by Eq. (3).

$$\alpha_{ij} = \frac{\text{area}(v_i) \cap \text{area}(v_j)}{\min(\text{area}(v_i), \text{area}(v_j))}$$

(area(v_i) is the fraction of the image area occupied by the $i^{th}$ bounding box. The intersection of the two bounding boxes is normalised by the smaller of the bounding boxes to ensure an overlap score of one, when a smaller object is inside a larger one. Inclusion of area(v_i) as a global edge feature causes the graph matching algorithm to match nodes of similar sizes to each other.

$\mu_{ig}$ and $\theta_{ig}$ are the magnitude and orientation of the edge connecting the centroid of the object corresponding to node $i$ to the global centroid. The global centroid is computed as given in Eq. (4).

$$c_g = \frac{1}{N} \sum_{k=1}^{N} c_k$$

$c_k$ represents the centroid of the $k^{th}$ local node. The global centroid represents the centre of the geometrical layout of the objects in the image. The edges connecting each object to the global node illustrate the placement of that object with respect to the overall object topology.

The proposed algorithm to construct a graph for a given image is presented in Algorithm 1. An example graph for an image is shown in Fig. 4.

4.3. Relative Importance of different objects

As discussed in Sec. 3, humans do not attach equal importance to all objects. The presence/absence of relatively smaller objects such as food carton, mobile and remote do not seem to make much of a difference in our perception of the images in Fig. 2. However, those images would have seemed significantly different to us, had the relatively larger objects (person/dog) been missing. This indicates varying importance for objects in the scene depending on their attributes. Further, Proulx [28, 29] perform visual search experiments to conclude that larger objects capture greater visual attention.

We deal with this unequal importance of different objects, by assigning a relative weight to each local node of
Algorithm 2: Attribute-Graph Ranking

**Input**: Query Image (Q)

**Output**: Ranklist

/* Construct query Attribute-graph using Algo. (1) */
[qGraph, bounding boxes \((R_1, R_2, \ldots, R_N)\), object classes \((o_1, o_2, \ldots, o_N)\)] = constructGraph(Q);

/* Weights indicate the relative importance of different query objects */

wts = getWeight\((R_1, R_2, \ldots, R_N)\);

/* Weights are then performed on the association graph, with inter-
mediate re-weighting jumps, while enforcing the matching
constraints. This graph matching algorithm has been shown
to be robust to noise, outliers and deformation. However,
our Attribute-Graph representation is general and any other
graph matching technique, capable of handling constraints,
can also be used to obtain matching scores between two
Attribute-Graphs. */

for image in Dataset images do

/* load Dataset image graph */
dGraph = loadGraph(image);

\([S_{ldl}, S_{gbl}, S_{edge}] = graphMatch(qGraph, dGraph,
\text{wts}, (o_1, o_2, \ldots, o_N));

/* Calculate score */

score = \(\alpha \times S_{ldl} + \beta \times S_{gbl} + (1 - \alpha - \beta) \times S_{edge}\)

Ranklist = sort(score);

An overview of the proposed method is depicted in Fig. 3.

4.4. Graph Matching

During the evaluation phase, we use graph matching to
compare the query image Attribute-Graph with each of the
dataset image Attribute-Graphs. Graph matching is done
such that the overall similarity score between the mapped
nodes and edges is maximised, while preserving the matching
constraints. We enforce the following constraints on the
matching process:

- A local node of a particular object class can be
  matched only to another local node of the same class.
- A global node can be matched only to another global
  node.

For our experiments, we use Re-weighted Random
Walks (RRWM) algorithm proposed by Cho et al. [2],
which formulates the problem of matching as the task of
node selection on an association graph. This graph is con-
structed by modelling the nodes as the candidate correspon-
dences between the graphs to be matched. Random walks
are then performed on the association graph, with inter-
mediate re-weighting jumps, while enforcing the matching
constraints. This graph matching algorithm has been shown
to be robust to noise, outliers and deformation. However,
our Attribute-Graph representation is general and any other
graph matching technique, capable of handling constraints,
can also be used to obtain matching scores between two
Attribute-Graphs.

The scores of each dataset image are given by Eq. (5).

\[ [S_{ldl}, S_{gbl}, S_{edge}] = GM(Q, D_k, \text{wts}, \text{objs}) \] (5)

where \(Q\) refers to the query image, \(D_k\) is the \(k^{th}\) image in
the dataset \(D\) and \(GM(.,.)\) is the graph matching technique.
\(S_{ldl}, S_{gbl}, S_{edge}\) are the scores associated with the match-
ing between the local nodes, the global node and the edges
respectively of the two graphs. \(\text{wts}\) correspond to the impor-
tance weights calculated as described in Sec. 4.3. \(\text{objs}\)
refers to the object classes associated with the local nodes.

The algorithm for obtaining a ranked set of images given
a query image is described in Algorithm 2. The final score
corresponding to a dataset image is calculated as shown in
algorithm. \(\alpha\) and \(\beta\) are constants determined empiri-
cally. The scores obtained are sorted to get the ranklist.

5. Datasets

To evaluate our ranking technique, we require datasets
which have graded relevance scores for the reference im-
ages corresponding to each query. Moreover, it is essential
that these relevance scores have been assigned considering
overall visual similarity, and not just one particular com-
ponent of the image (For e.g. according to ImageNet hi-
erarchy which focuses mostly on objects). For this purpose
we have created two datasets, \(rPascal\) and \(rImageNet\) which
are subsets of the aPascal [8] and ImageNet [5] databases.
Each of these datasets contain 50 query images and a set of
reference images corresponding to each query. Reference
images correspond to the set of images that are considered
for ranking for a given query image. These reference im-
ages have been selected so as to contain at least one object
in common with that particular query. This has been done
to prevent burdening the annotators with a lot of irrelevant
images.

12 volunteers were recruited to annotate the images.
Given a query image and the corresponding reference im-
ages, the annotators were asked to give a relevance score
between 0-3 for each of the reference images with respect
to that particular query. These score values were defined as
follows:

- 0 - irrelevant : Unrelated to the query and should not
  be retrieved when searching for this particular query
- 1 - fair : Has a few components/aspects similar to the
  query
- 2 - good : Similar to the query except for a few missing
  components/aspects
- 3 - excellent : Very similar, exactly what we would like
to see when searching for this particular query

Each annotator was presented with reference images cor-
responding to 20 query images from each dataset. This re-
sulted in the reference set corresponding to each query im-
age receiving annotations from at least 5 different annota-
tors. The final relevance score was obtained by computing
the median score of the 5 annotations for each image.
5.1. rPascal Dataset

The rPascal dataset [8] contains a total of 4340 images, spanning 20 labelled object classes, divided into 2113 training images and 2227 test images. Our rPascal (ranking Pascal) dataset is constructed entirely over the test set of aPascal. The queries consist of 18 indoor scenes and 32 outdoor scenes. The queries include both simple images containing one or two objects and also more complex images containing 8 or more objects. The rPascal dataset contains a total of 1,835 images with an average of 180 reference images per query.

5.2. rImageNet

The rImageNet dataset is constructed from the validation set of ILSVRC 2013 [30] detection challenge, which contains 20,121 images with objects belonging to 200 different classes. To increase the complexity of the dataset, we have selected only images containing at least 4 objects. The queries contain 14 indoor and 36 outdoor scenes. The rImageNet dataset contains a total of 3,354 images with an average of 305 reference images per query. Owing to the increased number of classes and objects in the images, this dataset is much more challenging than rPascal.

The statistics pertaining to the number of queries containing a certain number of objects for both the datasets is given in Fig. 5. Example queries from both datasets can be seen in the first column of Fig. 8.

6. Experiments and Results

6.1. Experimental setup

We perform object detection and classification, using Regions with CNN (RCNN), the algorithm of Girshick et al. [10]. These detected regions form the nodes of our graph and are characterised by the local attributes extracted from them. We employ the 64 attributes defined by Farhadi et al. [8] as our local attributes. They consist of shape attributes such as ‘2D boxy’, ‘cylindrical’ etc., part attributes such as ‘has head’, ‘has leg’, ‘has wheel’ etc. and material attributes such as ‘has wood’ and ‘is furry’. We use the same set of features as those used by Farhadi et al. [8], which describe each object by a 9751 dimensional feature vector containing texton, HOG, edge and colour descriptor based visual words, to train the attribute classifiers. The attribute classifiers are trained on the aPascal training set. For global attributes, we use the 205 dimensional output probabilities of a deep network trained on the Places database1[44], which have classes such as coast, desert, forest, home, hotel etc. We set $\alpha = 0.4$, $\beta = 0.4$ (in Algo. 2) in all our experiments.

For a baseline, we compare with our implementation of the work of Douze et al. [6]. Douze et al. uses Fisher vectors [27] (of SIFT features) extracted from the image concatenated with ‘Classemes’ [36] as features to describe an image. Classemes is a 2659 dimensional vector of classifier outputs, trained for categories that are selected from an ontology of visual concepts. The features used to obtain the classemes are extracted from the image as a whole and consist of colour GIST, oriented and unoriented PHOG, pyramid self similarity [33] and bag of words using the SIFT descriptor.

We also compare the proposed method with the attribute based ranking method of Cao et al. [1]. They represent an image using a set of disjoint triangles, whose vertices correspond to image objects. With this representation, they calculate the similarity between two images by mapping the triangles of one image on to the other.

6.2. Evaluation measures

For quantitative evaluation, we compute the Normalised Discounted Cumulative Gain (nDCG) of our ranklist. nDCG is a standard measure for evaluating ranking algorithms [1, 35] and is given in Eq. (6).

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i)}$$

$$nDCG@k = \frac{DCG}{IDCG} \quad (6)$$

1The pre-trained network was obtained at http://places.csail.mit.edu/
2Our implementation achieves the accuracies mentioned in [6] on the databases used by Douze et al.
Figure 6. Qualitative results: nDCG comparison against Ranking truncation level ‘k’ for (a) rPascal dataset (b) rImageNet dataset. The proposed method shows a significant improvement in performance over both Douze et al. [6] and Augmented Retrieval [1] as well as the other baselines in both datasets. Plots are best viewed by zooming in.

$n_{DCG}$ is the relevance of the $i^{th}$ ranked image and IDCG refers to the ideal Discounted Cumulative Gain(DCG), which acts as a normalisation constant to ensure that the correct ranking results in an nDCG score of 1. $k$ denotes the ranking truncation level.

6.3. Discussion

Figure 6 plots the nDCG scores of the proposed method, Douze et al. [6], Cao et al. [1], as well as a some other baselines as a function of the ranking truncation level $k$. The baselines we consider include nearest neighbour comparison with the fc7 features of AlexNet [16], PlacesNet [44], and a combination of the two. We also compare against a simple baseline which ranks images based on just the number of common objects present in the two images. As observed from the figure, our performance exceeds that of the other techniques [1, 6] as well as the above established baselines on both datasets. Since Douze et al. consider only the global descriptors of images and do not try to model the objects and their spatial layout, their results in many cases, fail to have all the desired objects in the first few retrieved images.

Cao et al. not only ignore the global scene context, but also do not characterise the relationships existing between their object triangles, thereby limiting the amount of spatial layout information up to three objects. As can be observed from Fig. 6, these drawbacks lead to a poor performance, especially when the queries are complex images with many objects as in the case of the rImageNet dataset.

On rPascal we show an improvement of 5-8% over Cao et al. and 7-12% over Douze et al. On rImageNet we show an improvement of 9-11% over Cao et al. and 8% over Douze et al.

Figure 8 shows a few ranking examples by the proposed method. The images ranked 1-7 for each of the queries in the first column are shown in columns 2-8 of Fig. 8. Ranked images annotated as fair and irrelevant have been marked with orange and red boxes respectively. Two of these appear in queries corresponding to rows D and H. This mis-ranking is mainly due to incorrect detections and classifications of objects in the reference images. For example, the two bikes in row D, column 6, are wrongly classified as cars. Similarly, the 6th ranked image in the last row has mis-detections of microphone.

The proposed method scales linearly with the number of images in the database. An inverted index file scheme based on objects and attributes can be used to speed this up further.

6.4. Ablation Analysis

To better understand the contribution of various components of the Attribute-Graph, we analysed the performance of the proposed method, by ablating each component. The results obtained are depicted in Fig. 7. Our analysis confirms that removal of any of the global node, object nodes, edges or importance weights assigned to the nodes, negatively impacts performance, thereby demonstrating the importance of each of the components of the proposed Attribute-Graph. We observe that the performance is most affected by the removal of object nodes leading to 7% and 12% drops in rPascal and rImageNet respectively.

7. Conclusion

We have proposed a novel image representation using an attribute based graph structure. Our Attribute-Graphs represent various objects present in the image, along with their characteristics such as shape, texture, material and appearance. Attribute-Graphs also capture the spatial scene structure via the graph edges and the overall scene gist through global attributes. We show the efficacy of the proposed representation by its application to image ranking. We evaluate the performance of our ranking technique on the rPas-
Figure 7. Ablation analysis of the proposed method for (a) rPascal (b) rImageNet. Plots are best viewed by zooming in.

Figure 8. Ranking results for the Proposed method: Column 1: Queries A-H, Columns 2-8: First 7 retrieved images. The first 4 queries (A-D) are from rPascal and the last 4 (E-H) belong to rImageNet. The annotation scores are shown at the right top corner of every image. Irrelevant images have been marked with a red boundary and images with an annotation of Fair have been marked with orange. The remaining images have been annotated as Good or Excellent.

cal and rImageNet datasets, which we have collated for the purpose. The proposed method obtains an improvement of around 5-8% in the nDCG scores, demonstrating the ability of our representation in capturing semantic similarity.

8. Acknowledgements

This work was supported by Defence Research and Development Organization (DRDO), Government of India. Supported by Microsoft Research India Travel Grant.
References

[40] H. Xu, J. Wang, X.-S. Hua, and S. Li. Image search by concept map. In *SIGIR*, 2010. 2