ISSN 2395-1621

Ranking Of Images Using Semantic Similarity With Duplicate Image Detection

[#1]Pooja Ombale, [#2]Pooja Indalkar, [#3]Aparna Gadade, [#4]Akshay Murme

[1]poojaombale9436@gmail.com
[2]poojaindalkar9458@gmail.com
[3]aparnagadade28@gmail.com
[4]akshaymurme143@gmail.com

[#1234]Department of Computer Engineering, Flora Institute of Technology,Savitribai Phule Pune University, Maharashtra, India

ABSTRACT

Image search is a particular data search used to determine images. For searching images, a user may give query terms such as keyword, image file, or click on few image, and the system will determine images "similar" to the query. The resemblance used for search criteria could be Meta tags, colour distribution in images, region/shape attributes, etc. Many Commercial Internet scale image search engines use only keywords as queries. The search engines (e.g. Google Image Search, Bing Image Search) mostly depend on surrounding text features. It is not easy for them to understand user's search intention only by query keywords and this leads to uncertain and noisy search results. It is important to use visual Information in order to solve the ambiguity in text-based image retrieval. In this paper, we propose a novel Internet image search approach. The user needs to click on one query image with the minimum attempt and images from a pool retrieved by text-based search are re-ranked based on both visual and textual content. To capture the user's search intention from this one-click query image in four steps has been presented in this paper. The proposed method uses a visual vocabulary of vector quantized local feature descriptors (SIFT) to find similarity measures to evaluate near duplicate image detection. Using this duplicate image detection technique with existing system improves precision of top ranked images as result demonstrates.

ARTICLE INFO

Article History

Received :15th April 2016

Received in revised form :

17th April 2016

Accepted : 19th April 2016

Published online :

23rd April 2016

Keywords— Image Search, Intention, Visual, Weight Schema, Clustering, Keyword Expansion, Duplicate Detection.

I. INTRODUCTION

Today's commercial Internet scale image search engines use only text information. Users type keywords in the hope of finding a certain type of images. The search engine returns thousands of images ranked by the text keywords extracted from the surrounding text. However, many of returned images are noisy, disorganized, or irrelevant. Even the state-of-the-art, such as Google Image Search and Microsoft Live Image Search , use no visual information. Using visual information to re-rank and improve text based image search results is a natural idea. Most of the existing works assume that there is one dominant cluster of images inside each image set returned by a keyword query, and treat images inside this cluster as "good" ones. Typical works include using each set of images returned by a keyword search to train a latent topic model or emphasize images that occur frequently. Unfortunately, all these approaches require online training, so cannot be used for real time online image search. In addition, these approaches cannot handle ambiguity inside a keyword query, since the assumption that images returned by querying one keyword is all from one class does not hold, and the structure of the returned image set is much more complicated. For example, the query for

© 2015, IERJ All Rights Reserved

"apple" can return images from 3 main classes such as Fruit apple, apple pie, and Apple digital products. Within each main class, there can be several distinct sub classes (images that are visually similar). Also, there are images that can be labelled as noise (irrelevant images) or neglect. In which a picture indexing and abstraction approach for pictorial database retrieval is used.

In this paper, we propose a framework and build a system to re rank text based image search results in an interactive manner. After query by keyword, user can click on one image, indicating this is the query image. We then re-rank all the returned images according to their similarities with the query. The most challenging problem in this framework is how to define similarity. There are many features in vision and CBIR community, either low level ones such as colour, texture, and shape, or higher level ones such as face. Using different features will produce different results, and there is no single feature that can work well for all images. How to integrate various visual features to make a decision about similarity between the query image and other images becomes the essential problem, especially for the diverse and open data set on the Internet. In CBIR community, the relevance feedback approaches focus on how to find a better combination weight of features based on multiple labelled images provided during multiple user feedback sessions. The performance has been limited. In addition, most approaches require online training based on the feedback samples, thus are difficult to be used for real time online applications.

II. LITERATURE SURVEY

Several works have done in web image re-ranking. some works have addressed the visual re-ranking process in visual search process. The below table gives an overview of all methods used in the visual search process with their merits and demerits. As the Table I shows existing technique can be categories into Classification based, Clustering based and Graph based methods. These methods use different technique for image re-ranking.

In the first row, classification based methods uses the pseudo relevance feedback is often utilized. Pseudo relevance feedback is a method that began from text retrieval. It takes few of the top ranked documents from the search results done initially as pseudo positive. Then in this method the different classifiers such as SVM, boosting and ranking SVM can be adopted. These techniques are very effective in data retrieval but it needs sufficient training. A lot of parameters need to be estimated.

The second method is clustering based method in which each sample is given a soft pseudo label according to the initial text search result, and then the Information Bottleneck Principle, Agglomerative Information Bottleneck is used for re-ranking, These method achieves good performance on the named-person queries. But it is limited to those queries which have significant duplicate characteristics.

The third method graph based method in which graph is constructed with the samples as the nodes and the edges between them being weighted by visual similarity. In this method Bayesian Visual Re ranking, Point wise ranking distance technique are used to re ranked the images which are very effective also increase visual consistency and reduce ranking distance. But it fails to capture disagreement between score list.

Web mining is important aspect for users to get the data to be highly accurate and most relevant to user query and to at most satisfaction of user. Web Image Re-ranking is an effective way to improve the results of web-based image search. A major challenge in web image Re-ranking is that the similarities of visual features do not well correlate with images and semantic meanings which cannot interpret users search intention. By providing keyword expansion to textual query and visual query expansion to image query the results retrieved are promising than previous implemented systems.

TABLE I. LITERATURE SURVEY

Categories	**Merits**	**Demerits**
Classification Based Method	Very effective in data retrieval	Sufficient training needed, lot of parameter needed, complexity in Designing.
Clustering Based Method	Good performance on name person queries	Limited to those queries which have significant duplicate characteristics.
Graph Based Methods	Very effective, Increase visual Consistency and Reduce ranking distance.	Fails to capture disagreement between score list.

III. EXISTING SYSTEM

In existing system, one way is text-based keyword expansion, making the textual description of the query more detailed. Existing linguistically-related methods find either synonyms or other linguistic-related words from thesaurus, or find words frequently co-occurring with the query keywords. Google image search provides the Related Searches feature to suggest likely keyword expansions. However, even with the same query keywords, the intension of user can be highly diverse and cannot be accurately capture by these expansion. Search by image is optimized to work well for content that is reasonably well described on the web. For this reason, you will likely get more relevant result for famous landmarks or paintings than you will for

© 2015, IERJ All Rights Reserved

more personal images like you toddler's latest figure painting.

The existing system also detects duplicate images and removed by comparing hash codes. World Wide Web contains billions of images. User browsing the internet will quickly encounter duplicate images in multiple locations. Duplicate image detection is done for reducing storage space, understanding behaviour and interest of user and for copyrights. Duplicates can be exact duplicates, global duplicates or near duplicates. Exact duplicate images have exactly the same appearance that is images with identical contents. The small alterations in the content of image are ignored in global duplicates. Near duplicate allows rotation, cropping, transforming, adding, deleting and altering image content. In traditional duplicate image detection system, images are first converted into a particular image representation and then stored in indexing structure.
When query image is received, system uses the indexing structure and similarities are computed by assigning score to each candidate image based on query image. Then certain threshold is applied to determine which of the candidate image are truly duplicates of the query image. The main contribution of this paper is to present a modified similarity measure which improves performance of duplicate image detection.

IV. PROPOSED SYSTEM

In this section, we will describe a easy and efficient image re-ranking system. The system Architecture contains the mainly two parts that is online part and offline part. The figure can show. The online part in that search the image on the bases of text. Some text can be enter in the search engine. The images can be present on the search result with the help of re ranking of query images. If you enter query as a 'mouse' that time all mouse related images can be view by user. The keyword related references classes can be retrieved. The duplicated information can be return to the database. The following modules can be work on system that are as follows:

1] Image search: Image search is nothing but a data search used to find data in search engine.

2] Query expansion: We can enter any web query keyword in search engine on web page then different result are categorize in different form.

3] Visual query expansion: when user search any query into to search engine that time some images are provided for user .On the basis of visual feature image can re-rank. The expansion means the small database can be created and its related all sub information can be expand.

4] Image retrieved by keyword expansion: If the user can enter some query on search engine that time in the database that's related information can be search and get the result to user requirements.

A. **The K-Means Algorithm Process:**

The K-means algorithm can be implemented on these systems. The K-means algorithm is the simplest clustering algorithm. It depends on the unsupervised learning algorithm. The unsupervised algorithm means no training set are available. The procedure is to follow the simple and easy structure in that making a one cluster. The data sets contain the multiple data point and the K-means algorithm can handle the linear data only. The consider data point is x={ x1, x2,.....xn } .find follows the K-means algorithm.

1] Randomly select 'c' cluster center

2] Calculate the distance between each data point and cluster center

3] Assign the data point to the cluster center whose distance from the cluster is minimum of all the cluster center
4] Recalculate the cluster center using cluster formula

5] Recalculate the distance between each data point and new obtained cluster center.

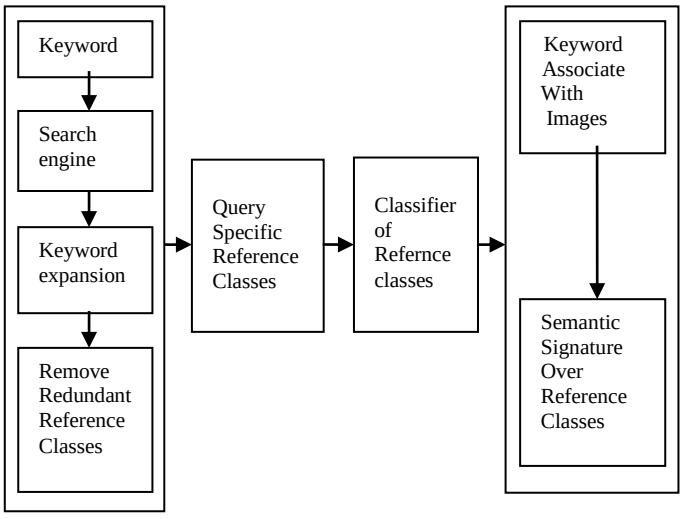

FIGURE I. PROPOSED SYSTEM

V. SYSTEM ARCHITECTURE

First User enters a keyword on the search engine, which retrieves the thousands of images, from which user clicks on one query image. The query image is first categorized into adaptive weight categories. Under every category, a specific pre-trained weight schema is used to combine visual features adapting to this kind of images to improved re-rank the text-based search result. Re-rank images based on keyword based expansion is performed using tfidf *(*term frequency–inverse document frequency*)* method. A word w is recommended as an expansion of the query if a cluster of

© 2015, IERJ All Rights Reserved

images are visually similar to the query image and all contain the same word w. The expanded keywords better capture users' search intention since the consistency of both visual content and textual description is ensured. Perform clustering of images based on the visual and textual similarity, a cluster of images visually similar to query image are found, which are used as multiple positive image examples from which textual and visual similarity metrics is obtained. These metrics used for image re ranking, because they are more specific and robust to the query image.

Efficient re-ranking of images is done using visual and textual similarity. Finally detection of duplicate images based on similarity measures by using duplicate image detection algorithm is done. In this algorithm, first images are matched using distinctive invariant features. These features are extracted from set of reference images using Scale Invariant Feature Transform (SIFT) algorithm and stored in database. A new image is matched by comparing each feature of new image to this previous database.

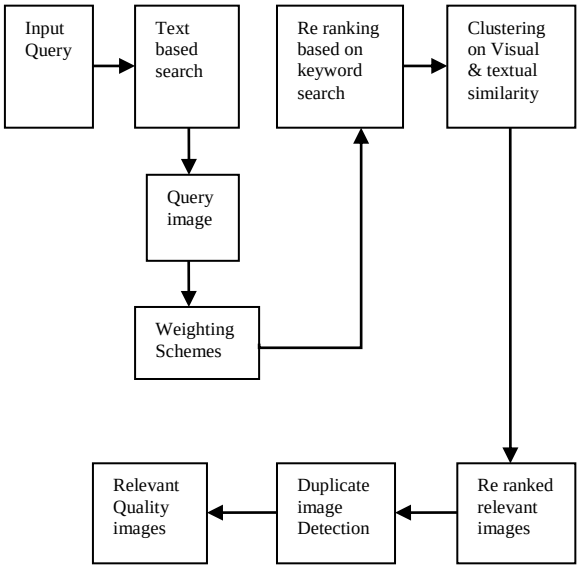

FIGURE II. SYSTEM ARCHITECTURE

VI.RESULTS

Quantifying the results of an image retrieval system is not a trivial task. While most CBIR papers use a hand labeled corpus of images to compute precision and recall on a set of queries, this is not appropriate for the task of refining the results of a web-based image search engine. To competently perform evaluation of such a system, a study must be made using human subjects who would evaluate the qualitative performance. In this study we concentrated on the algorithmic aspects of discovering a visual representation of a given query, and we leave such evaluation for future work. It is important to note that if a query is vague, or has multiple semantic meanings; our method will favor the more prominent meaning. The detection of multiple meanings per query seems plausible, but in this paper we chose to concentrate on single meanings, with the argument that a query could be made more specific. Another important note about any system similar to what we have described in this paper, is that it must be efficient in order to be used in a real application. In our implementation, the segmentation and feature extraction could be done offline on an entire database .The rest of the processing takes a few seconds, which is fast enough for a real-time application.

In fig 3.the image keyword relationship chart represents the comparison of total image versus image keyword. It can be plotted after the bag annotation process. Here it can display the total number of relevant and irrelevant images to the user. In this fig apple keyword contains 13 relevant images. In keyword graph contains 2 relevant and 1 irrelevant images

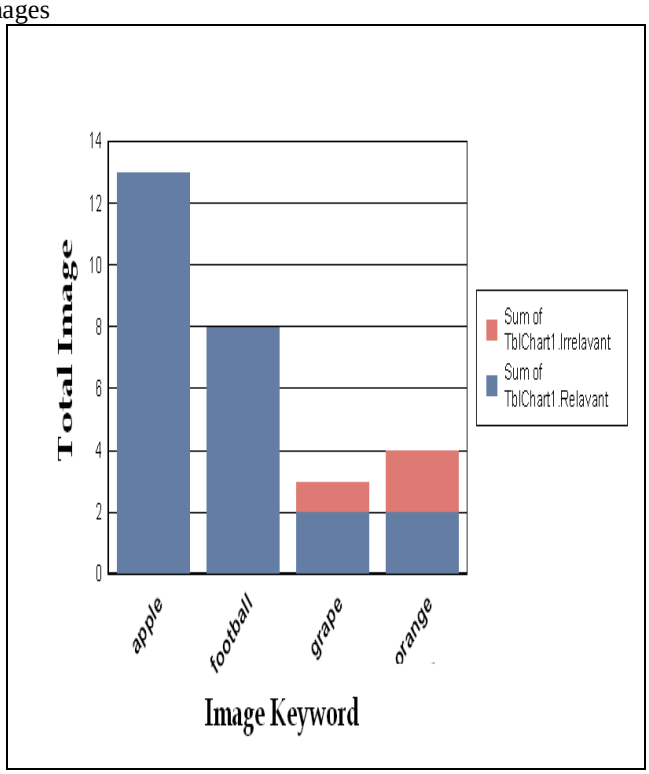

FIGURE III. KEYWORD RELATIONSHIP CHART

In fig.4.the positive image ranking score can be displayed for every keyword. It shows the ranking percentage of comparison result with other images. Here every images ranking percentage can be varied according to the relevant images in the positive bag. The higher percentage of ranking score has the more relevant images which moves to the positive bag whereas below 70 percentage moves to the negative bag which contain irrelevant images. This provides an more efficient re-ranking of images compare to the existing method.

.

© 2015, IERJ All Rights Reserved

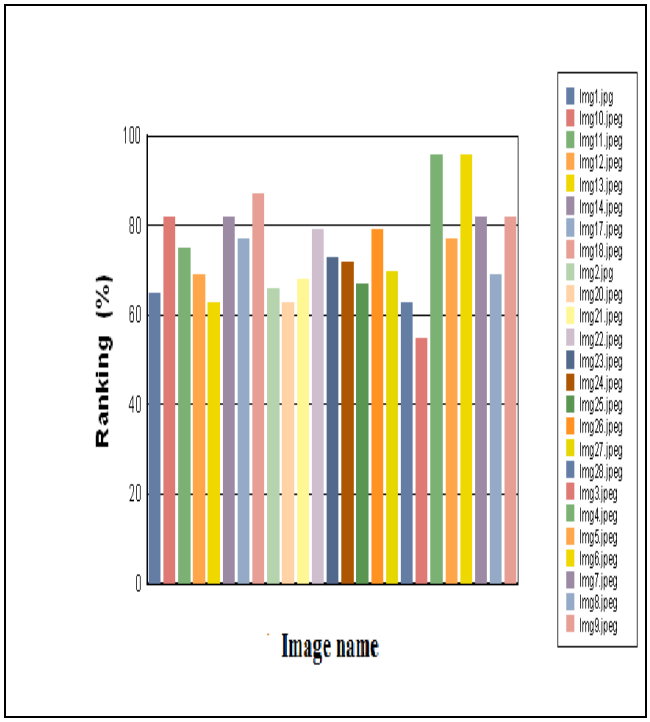

FIGURE IV. IMAGE RANKING SCORE CHART

VII. CONCLUSION

Image search is a particular data search used to find images. To search for images, a user may give query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query. In this, a novel Internet image search approach which only needs one-click user response. Intention specific weight schema is proposed to combine visual features and to compute visual similarity adaptive to query images. Without additional human feedback, textual and visual expansions are integrated to capture user intention. Expanded keywords are used to extend positive example images and also enlarge the image pool to include more relevant images.This framework makes it possible for industrial scale image search by both text and visual content. The proposed new image re-ranking framework consists of multiple steps, which can be improved separately or replaced by other techniques equivalently effective. Our system will overcome the drawbacks of existing system by generating quality and exact match result of user intention and the additional function stops retrieving duplicate images and also the repeated images are detected and avoided by system in output. So user will be getting final output as plain, intended images.

ACKNOWLEDGEMENT

We are very grateful to all authors in reference section. Their methods, algorithms, conceptual techniques are very helpful for our research. All papers in the reference section are very useful for our proposed system.

REFERENCES

[1] Xiaoou Tang1,Ke Liu2, Jingyu Cui3,Fang Wen4 and Xiao gang Wang5 “Intent Search: Capturing User Intention for One-Click Internet Image Search” 2012.

[2] J. Deng1, A.C. Berg2 and L. Fei-Fei3 “Hierarchical Semantic Indexing for Large Scale Image Retrieval”, in Proceedings of International Conference Computer Vision and Pattern Recognition, 2011.

[3] J. Krapac1, M. Allan2, J. Verbeek3 and F. Jurie4 “ Improving web image search results using query-relative classifiers,”□in Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2010 .

[4] G. Chechik1, V. Sharma2, U. Shalit3 and S. Bengio4, “Large Scale Online Learning of Image Similarity through Ranking”, J. achine Learning Research, vol. 11, 2010.

[5] Z. Zha1, L. Yang2, T. Mei3, M. Wang4 and Z. Wang5, “Visual Query Expansion”, in Proceedings of IEEE 17th ACM International Conference on Multimedia, 2009.

[6] J. Cui1, F. Wen2, and X. Tang3 ” Real Time Google and Live Image Search Re-Ranking”, in Proceedings Of IEEE- 16th AMC International Conference on Multimedia, 2008.

[7] Jingyu Cui1, Fang Wen2, Xiaoou Tang3, “Intent Search Interactive On-line Image Search Re-ranking”,2008.

[8] F. Jing1, C. Wang2, Y. Yao3, K. Deng4 L. Zhang5 and W.Ma6, “Igroup: Web Image Search Results Clustering”, in Proceedings of IEEE- 14th Ann ACM International Conference on Multimedia, 2006

[9] N. Ben-Haim1, B. Babenko2, and S. Belongie3 Improving Web- Based Image Search via Content based clustering”, in Proceedings of International Workshop Semantic Learning Applications in Multimedia, 2006 .

[10] K. Tieu1 and P. Viola2 “Boosting Image Retrieval” International J. Computer Vision, 2004.

© 2015, IERJ All Rights Reserved