Bridging the Semantic Gap in Image Search via Visual Semantic Descriptors by Integrating Text and Visual Features

V.L. Lekshmi and Ansamma John

Abstract To facilitate access to the enormous and ever–growing amount of images on the web, existing Image Search engines use different image re-ranking methods to improve the quality of image search. Existing search engines retrieve results based on the keyword provided by the user. A major challenge is that, only using the query keyword one cannot correlate the similarities of low level visual features with image's high-level semantic meanings which induce a semantic gap. The proposed image re-ranking method identifies the visual semantic descriptors associated with different images and then images are re-ranked by comparing their semantic descriptors. Another limitation of the current systems is that sometimes duplicate images show up as similar images which reduce the search diversity. The proposed work overcomes this limitation through the usage of perceptual hashing. Better results have been obtained for image re-ranking on a real-world image dataset collected from a commercial search engine.

Keywords Image re-ranking · Visual semantic descriptor · Semantic space · Perceptual hashing · Image search

1 Introduction

Web image search and retrieval has become an increasingly important research topic due to abundance in multimedia data on internet. Most of the web scale image search engines mainly use two schemes for searching for images on the web. In the first, keyword based scheme, images are searched for by a query in the form of

V.L. Lekshmi (✉) · A. John
Department of Computer Science and Engineering, TKM College of Engineering, Kollam, Kerala, India
e-mail: lekshmivl20@gmail.com

A. John
e-mail: ansamma.john@gmail.com

M. Senthilkumar et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 412, DOI 10.1007/978-981-10-0251-9_21

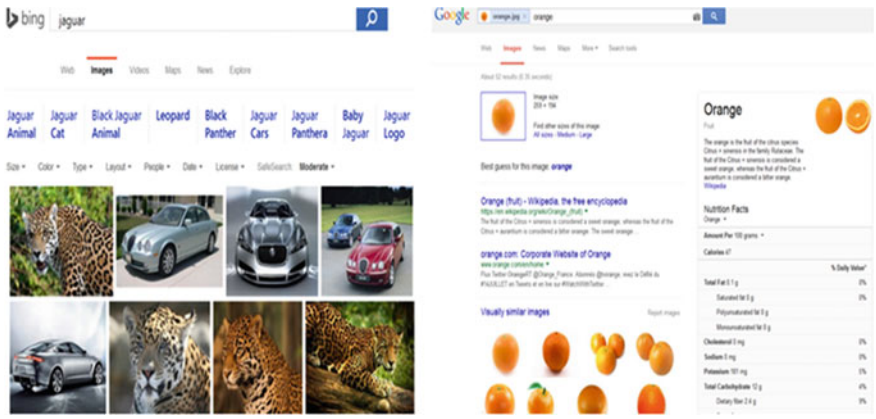

Fig. 1 Illustration of the keyword based and example based image search scheme

textual keyword provided by the user. The second, example based scheme allows the users to search for similar images by providing an example serving as query image both schemes are illustrated in Fig. 1.

Independent of which search scheme is deployed, an image search engine generally operates in two main steps: the offline and the online step. For many query keywords the image retrieval performance is good, but the precision of the returned results is still relatively low. They suffer from the ambiguity of query keywords, because it is difficult for users to accurately describe the visual content of target images only using query keywords. One of the major challenges is the conflict between the content of the image and the webpage textual information. This paper attempts to resolve this bottleneck by depending on both the textual information and visual information. Another major challenge in the existing systems is that its similarities of low level visual features may not correlate with image's high level semantic meanings. To reduce this semantic gap, visual features are mapped to a predefined attributes known as visual semantic descriptors.

In this paper we propose a web image search approach, which requires both query keyword and query image. First a text based search is performed by using a query keyword. From the pool of images retrieved, the user is asked to select the query image. Images in the pool are re-ranked based on the visual semantic descriptors of the query image. This query-specific visual semantic descriptor effectively reduces the gap between low-level visual features and semantic categories, and makes image matching more consistent with visual perception.

Another major issue in web image search is removing near-duplicate images. We address the diversity problem to make the search result more diverse and to improve efficiency of the search engine. Identifying distinct images can prevent duplicate images from being used once they are uploaded. In this paper we use Perceptual hash method to remove duplicate images and improve search diversity. Image features are used to accomplish this and experimental results show the better

improvement of diversity in the search result. Encouraging results are obtained for proposed image re-ranking method on a real world image dataset collected from commercial search engines.

The rest of this paper is organized as follows. We discuss related works in Sect. 2 and follow up with the details of our method in Sect. 3 and present evaluation results in Sect. 4 and finally conclusions and future works are offered in Sect. 5.

2 Related Work

Most of the internet scale image search engines like Google Image Search and Bing primarily depend on the textual information. The annotation error due to the subjectivity of human perception is one of the major disadvantage of this approach. Text based image search suffers from the ambiguity of query keywords [1]. Content based image retrieval (CBIR) was introduced in early 1980s, to overcome the limitations of text based image search. Visual features of images are used in CBIR to evaluate image similarity.

Generally there are three categories of visual re-ranking methods classification based, clustering based and graph based. Classification-based methods first select some pseudo-relevant samples from the initial search result. Deng [2] learned to produce a similarity score for retrieval by using a predefined comparison function based on a known hierarchical structure. In clustering based methods the pool of images in the initial search result are first grouped into different clusters. According to the cluster conditional probability, order the clusters and re-ranked result list is created. By using the cluster membership value order the samples within each cluster. Graph based methods are more recently proposed and have received increased attention. In this method first a graph is built with the images in the initial search result serving as nodes. If two images are visual neighbors of each other, an edge is defined between those images and the edges are weighted by the visual similarities between the images. For instance, either a random walk over the graph or an energy minimization problem, re-ranking can be formulated.

Cui [1, 3] classified query images into eight predefined adaptive weight categories, inside each category a specified pre-trained weight schema is used to combine visual features. For reducing the semantic gap, query specific semantic signatures was first proposed in [4]. The proposed visual semantic descriptor is effective in reducing the semantic gap when computing the similarities of images. One shortcoming of the existing system [1, 3, 5] is that sometimes duplicate images show up as similar images to the query. The proposed system overcomes this problem by adding a perceptual hash method to detect duplicate images. Our work incorporates both textual and visual information for re-ranking, moreover incorporation of textual information and duplicate detection significantly improves the search result and was not considered in previous work.

3 Method

We use both query keyword and query image for searching the images. The images for re-ranking are collected from different search engines like Google, Bing etc. User first submits a query keyword, from the pool of images returned after text based search, user is asked to select a query image which reflects the user's search intention. Based on the visual similarity with the query image, images in the pool are re-ranked. The semantic classes associated with query keywords are discovered by expanding the query keywords from the result obtained from the text based image search by utilizing both text and visual features. This expanded keyword defines the semantic classes for the query keywords. For each semantic class the training images are collected. A multilayer perceptron on the text and visual features is trained from the training sets of the semantic classes and its output is stored as visual semantic descriptor which describe the visual content from different aspects.. We combine the described features together to train a single classifier, which extracts a single semantic descriptor for an image and the extracted descriptors are stored. These semantic descriptors are used to compute the similarities for re-ranking. It provides much better re-ranking accuracy, since training the classifiers of reference classes captures the mapping between the visual features and semantic meanings [5]. The Flowchart of our approach is shown in Fig. 2.

3.1 Feature Design

We adopt a set of features which describe the visual content of an image in different aspects. For reducing the semantic gap both high level features and low level features are used to capture the user's search intention. Here we briefly explain the features used in this work.

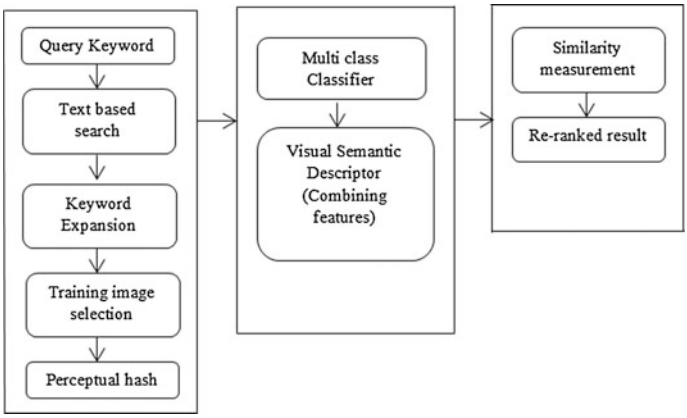

Fig. 2 Overall block diagram

- *Color*: The most widely used color feature used in image retrieval is color. One color description that is good to dispense efficiency and effectiveness in illustrating the distribution of colors of an image is color moment [6]—Mean, standard deviation, skewness. Mathematically those moments are defined as follows [7].
- *Texture*: Texture is one of the major feature used in recognizing objects [8]. Here we use co-occurrence matrix for image texture analysis.
- *Attention Guided Color Signature*: The color composition of an image describes color signature [9]. Clusters centers and their relative proportions are taken as the signature.
- *Histogram of Gradient (HoG)*: HoG reveals distributions of edges over different parts of an image, and is especially effective for images with strong long edges [10].
- *Wavelet*: We use the 2nd order moments of wavelet coefficients in various frequency bands to characterize the texture properties in the image [11].

3.2 Expanding the Keywords and Collection of Training Images

The keyword expansions most relevant to the query keyword are identified to define the semantic classes associated with each query keyword (For example, for a query keyword '*diamond*' some of the semantic classes identified are *black diamond, diamond jewelers, red diamond* etc.). To find the keyword expansions first a text based search is performed and the top ranked images are retrieved from the text based search. Keyword expansions are found from words extracted from the top retrieved images. The T most frequent words $W_1 = \left\{w_I^1, w_I^2, \ldots, w_I^T\right\}$ among top D re-ranked images are identified and these words are used for expanding the query keyword. The extracted words are sorted according to the frequency of words become available among the D images from higher to lower. If a word w is among the top ranked image, it has a ranking score $r_i(w)$ according to its ranking order. Otherwise ranking score must be zero [5].

$$r_I(w) = \begin{cases} T - j & w = w_I^j \\ 0 & w \notin W_I \end{cases} \tag{1}$$

To obtain the training images of semantic classes, we combine the words obtained from the keyword expansion with original query keyword and it is used as the query to the search engine. From the search result top K images are taken as the training images for the semantic classes. To reduce the computational cost remove the similar semantic classes this is both semantically and visually similar.

3.3 Perceptual Hash Method

To identify the duplicate images here we use perceptual hash method. Duplicate images significantly reduce the storage space and decrease the search diversity. Perceptual image hash functions create hash values based on the image's visual appearance. This function computes similar hash values for similar images, whereas for different images dissimilar hash values are calculated. Finally, using a similarity function to compare two hash values, it decides whether two images are different or not. The different steps involved in the method are as follows:

- Reduce size: Method starts with a small image. 32 × 32 is a good size; this step is done to simplify the DCT (Discrete Cosine Transform) computation.
- Reduce color: The image is reduced to grayscale just to further simplify the number of computations.
- Compute the DCT: DCT disparate the image into a collection of frequencies and scalars, here uses a 32 × 32 DCT.
- Reduce the DCT: While the DCT is 32 × 32, to represent lowest frequencies in the image keep the top-left 8 × 8.
- Calculate the average value: Mean DCT value is computed.
- Further reduce the DCT: Depending on whether each of the 64 DCT values is above or below the average value set the 64 hash bits to 0 or 1.
- Construct the hash: Set 64 bits into a 64-bit integer. To understand what this fingerprint looks, set the values (based on whether the bits are 1 or 0 this uses +255 and −255) and convert from the 32 × 32 DCT (with zeros for high frequencies) back into the 32 × 32 image. Construct the hash from each image to compare two images, and count the number of bit positions that are different. (Hamming distance) A distance of zero indicates that it is very similar pictures thus identifies the duplicate images and eliminate it thereby improving the search results.

3.4 Combining Features and Visual Semantic Descriptor Extraction

The above described features characterize the images from different perspective of color, shape and texture. The objective of using a visual semantic descriptor is to capture the visual content of an image. Here we combine all the visual features to train a single multilayer perceptron better distinguishing reference classes. The output of the classifier is taken as the visual semantic descriptor. If S semantic classes for a query keyword, classifier on the visual features is trained and it outputs an S dimensional vector v, which indicates the probability of the image belonging to different semantic classes and it is stored as the visual semantic descriptor of the

image. The distance between two images I^a and I^b are measured as the $L1$ distance between their semantic descriptors v^a and v_b,

$$d(I^a, I^b) = \left\| v^a - v^b \right\|_1 \tag{2}$$

4 Experimental Results

From different search engines, the images are collected for testing the performance of re-ranking. Averaged top m precision is used as the evaluation criterion. Top m precision is defined as the proportion of relevant mages among top m re-ranked images. Averaged top m precision is obtained by averaging over all the query images. The averaged top m precisions on dataset are shown in Fig. 3. The improvements of averaged top 10 precisions on the 10 query keywords on dataset by comparing text and visual based methods are shown in Fig. 4. Here we choose 10 semantic classes for a single query keyword and only one classifier is trained combining all types of features, so the semantic descriptors are of 10 dimensions on average. For a particular semantic class fifty images are selected, so for a single

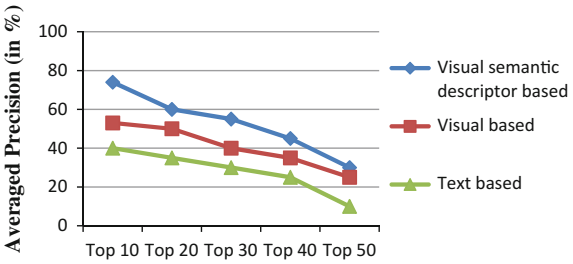

Fig. 3 Averaged top m precisions on different methods

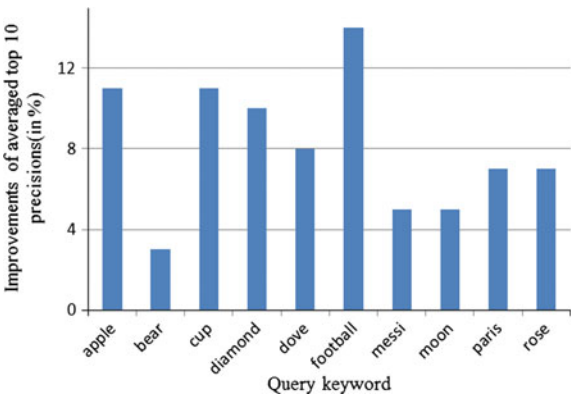

Fig. 4 Improvements of averaged top 10 precisions on the 10 query keywords

query keyword 500 images are used. For training, 75 % of the images were used and remaining 25 % were utilized for testing.

In this work we combine all the extracted features to train a single classifier to reduce the computational overhead caused when training separate classifiers for each feature. The dimensionality of extracted features is very large, so in the existing systems more computational cost is for comparing the visual features of images when re-ranking. Here we overcome this computational overhead by the use of visual semantic descriptors which have very less dimension compared to the visual features. The use of perceptual hashing method increases the diversity in search result by removing the duplicate images. The proposed work significantly outperforms the text based and visual based methods which directly compare visual features. The averaged top 10 precision is enhanced from 53 % (visual based) to 74 %. 21 % relative improvement is achieved.

5 Conclusion and Future Work

We propose a Web image search approach which requires both query keyword and query image. Previous methods only use either textual information or example input by the user. Visual semantic descriptors are proposed to combine the visual features and to compute the visual similarity with the query image. User intention is captured by both textual and visual information. The proposed image re-ranking framework incorporates duplicate detection to identify the duplicate images to the query image which make the search result more diverse. In the future we plan extend this paper for video event recognition [12]. To further improve the quality of re-ranked images, we aim to combine this work with photo quality assessment work in [13].

References

1. Tang, X., Liu, K., Cui, J., Wen, F., Wang, X.: Intentsearch: capturing user intention for one-click internet image search. IEEE Trans. PAMI **34**, 1342–1353 (2012)
2. Deng, J., Berg, A.C., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2011)
3. Cui, J., Wen, F., Tang, X.: Real time google and live image search re-ranking. In: Proceedings of the ACM Multimedia (2008)
4. Wang, X., Liu, K., Tang, X.: Query-specific visual semantic spaces for web image re-ranking. In: Proceedings of the CVPR (2010)
5. Wang, X., Qiu, S., Liu, K., Tang, X.: Web image re-ranking using query-specific semantic signatures. TPAMI (2013)
6. Stricker, M., Orengo, M.: Similarity of color images. In: IS&T and SPIE Storage and Retrieval of Image and Video Databases III, pp. 381–392 (1995)

7. Maheshwary, P., Sricastava, N.: Prototype system for retrieval of remote sensing images based on color moment and gray level co-occurrence matrix. IJCSI Int. J. Comput. Sci. Issues **3**, 20–23 (2009)
8. Haralick, R.M., Shanmugam, K., Its'Hak, D.: Textural features for image classification. IEEE Trans. Syst. Man Cybernetics **3**(6), 610–621 (1973)
9. Rubner, Y., Guibas, L., Tomasi, C.: The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In: Proceedings of the ARPA Image Understanding Workshop (1997)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2005)
11. Unser, M.: Texture classification and segmentation using wavelet frames. IEEE Trans. Image Process. **4**(11), 1549–1560 (1995)
12. Duan, L., Xu, D., Tsang, I.W., Luo, J.: Visual event recognition in videos by learning from web data. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1959–1966 (2010)
13. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2006)