International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Review on Content Based Duplicate Image Detection

**Jagtap Ankita K.[1], Tidke B. A[2]**

[1] Computer Networking Department Flora Institute of Technology, Pune- India

**Abstract:** ***Content Based image retrieval uses visual contents to search images from large scale database according to user's requirement. This paper provides a review on features used in content based image retrieval and various methods used to reduce semantic gap. It also introduced new system for retrieval of relevant images with duplicate detection framework. The proposed method uses a visual vocabulary of vector quantized local feature descriptors to find similarity measures to evaluate near duplicate image detection. Using this duplicate image detection technique with existing Intent search system improves precision of top ranked images.***

**Keywords:** Content- Based image retrieval, visual, semantic gap, duplicate image detection**.**

## 1. Introduction

The size of digital image collection is increasing rapidly, with the development of internet and availability of image capturing devices. To efficiently store and retrieve images for different domains many general purpose image retrieval systems have been developed. There are two approaches text-based and content-based for image retrieval. Text-based approach was invented in 1970's, in which images are manually annotated by text descriptors. To perform image retrieval database management system (DBMS) uses these descriptors. Text based approach have some disadvantages. The first is that image annotation requires much time and human labors. The second is inaccuracy in annotation due to different semantic views for each user. Third human intention and image contents are not taken into account. To overcome the difficulties in text-based approach content based image retrieval (CBIR) was introduced in the early1980s. CBIR uses visual contents to search images from large scale image database as per users' query. "Content based" means that the search will analyzes the actual contents either color, texture, shapes or spatial locations of the images. Chang published a pioneering work in 1984, in which a picture indexing and abstraction approach for pictorial database retrieval is used [1]. In CBIR low level features are extracted from the query image and these features are compared with features database. Images having least distance with query image are displayed as the result.

### 1.1 Low level Image Features

CBIR system is based on feature extraction, image features can be extracted globally for the entire image or locally for regions. There are two types of features, general features which are application independent and domain specific features which are application dependent such as human face, fingerprints and conceptual features [2].

#### 1.1.1 Color Feature

It is most widely used feature in CBIR system. Colors are defined on selected color space such as RGB, LAB, LUV, and HSB. Color feature or descriptors in CBIR system includes color moments, color histogram, invariant color histogram and dominant color. Color histogram is employed to represent the distribution of color of an image. It is graph which contains the occurrence of each intensity value found in that image, obtained by counting all image pixels having that intensity value. The number of bins of histogram determines the color quantization. The similarity between two histograms is computed by performing L1, L2 or weighted Euclidean distance or by computing their intersection [3]. In CBIR system images are not pre-processed, often corrupted with noise hence preprocessing is essential.

#### 1.1.2 Texture Feature

In image classification texture feature provides an important information as it describes the content of real-world images such as skin, clouds, trees, fabric etc. Texture represents structural arrangement of surface and their relationship to the surrounding environment. Texture features are of two types, first spectral features, such as features obtained using Gabor filtering or wavelet transform, second statistical features characterizing texture in terms of local statistical measures, such as the six Tamura texture features and Wold features. Coarseness, directionality, regularity, contrast, line-likeness, contrast and roughness are six tamura features [4].

#### 1.1.3 Shape Feature

Shape feature is important image feature. It includes aspect ratio, circularity, Fourier descriptors, moment invariants, Consecutive boundary segments etc. [5]. Shape representation can be divided in two categories, boundary based and region based. In boundary based representation only outer boundary of image is used. In region based uses the entire shape regions.

#### 1.1.4 Spatial Location Feature

Spatial location is also useful in region classification, instead of color and texture. For example sky usually appears at the top of an image and sea at the bottom, but both having similar color and texture feature. According to the location of the region in an image, spatial locations are defined as 'upper, bottom, top' [6]. A spatial context modeling algorithm is presented in Ref [7] which considers six spatial relationships between region pairs: left, right, up, down, touch and front to better support semantic based image retrieval.

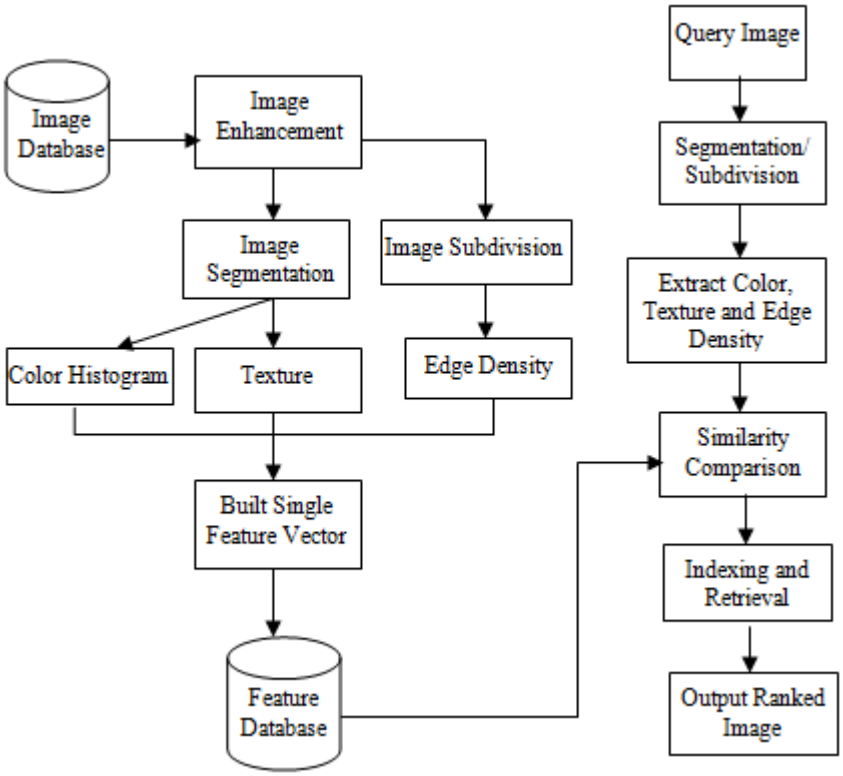


**Figure 1**: Flowchart of Content Based Image Retrieval [14]

### 1.2 Semantic Gap

The problem of CBIR system is the semantic gap. Semantic gap is referred as, the discrepancy between the limited descriptive power of low level image features and richness of user semantics. Techniques to reduce the semantic gap are divided into five categories, using object ontology, using machine learning tools, relevance feedback (RF), generating semantic template (ST) and making use of visual and textual information.

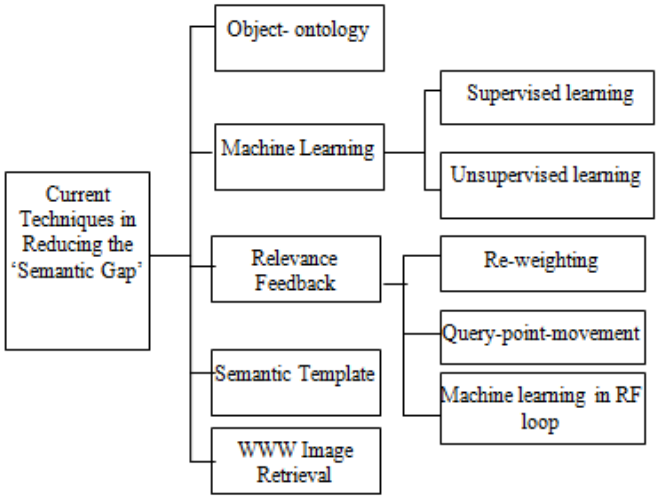


**Figure 2**: Current Techniques in Reducing Semantic Gap [4]

#### 1.2.1 Object Ontology

Semantic gap in image retrieval can be reduced using object ontology. Description of the image is nothing but the semantic of the image. The object ontology provides qualitative definition of the high level query concepts. Based on information, database images can be classified into different categories by mapping descriptors to high-level semantics (keywords) [8]. For example, sea can be defined as region of light blue color, uniform texture and bottom spatial location.

#### 1.2.2 Machine Learning

There are two types of tools in machine learning such as supervised in which value of outcome measure is predicted based on a set of input measure and unsupervised learning, in which the goal is to describe how input data is organized or clustered [9]. To learn high level concepts from low level image features Support Vector Machine (SVM) is widely used. SVM is used for object recognition, text classification, and binary classification [10]. Image clustering is unsupervised learning technique for retrieval purpose. Set of image data is grouped in such a way that the similarity within a cluster should be maximized and similarity between different cluster must be minimized [4].

#### 1.2.3 Relevance Feedback (RF)

RF tries to learn users' intention in on-line processing. RF reduces the semantic gap with help of user interaction and provides significant performance boost in CBIR systems [11]. In RF, with query-by-example, sketch etc, system provides initial retrieval results. Then user is asked to select relevant (positive examples) and irrelevant (negative examples) images to the query, then machine learning algorithm is applied to learn the users' feedback [12].

### 1.3 Performance Evaluation

CBIR system uses precision and recall to measure retrieval performance. Precision (Pr) is ratio of the number of relevant images retrieved (Nr) to the number of total retrieved images K. Recall is defined as ratio of number of retrieved relevant

images (Nr) over the total number of relevant images available in the database (Nt) [4] .

$$Pr = \frac{Nr}{K} \quad 1$$

$$Re = \frac{Nr}{Nt} \quad 2$$

### 1.4 Duplicate Image Detection

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 behavior 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 [14]. 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.

## 2. Existing Work

[14] Proposed a two step approach which combines local and global features. Seed clusters are discovered based on global descriptors with high precision and local descriptors used to grow the seeds to improve recall. Efficient hashing technique and MapReduce framework is used for duplicate image discovery. Navdav B.-Haim et.al.[15] Proposed ReSPEC (Re-ranking Set of Pictures by Exploiting Consistency) approach, combination of two methods. First step of algorithm retrieves the results by keyword query from an existing image search engine. In second step based on extracted image features, query results are clustered, which are most relevant to search query. In order of relevance remaining results are reranked. Image is segmented in blobs and features are extracted and clustered using mean Shift algorithm. Feature extraction is limited to color only, additional features would be necessary to handle more challenging problems. Li chen et. Al. [16] Proposed attention based similarity measure, which extracts colors and texture based signatures to compare near duplicate images. In pair of images similarity is determined by amount of matching structures detected. Color histogram intersection and Gabor based signature matching used to compare the proposed method. Ondrej chum et.al. [17] proposed two image similarity measures, which perform fast indexing based on locality sensitive hashing. Proposed approach uses a visual vocabulary of vector quantized local feature descriptors (SIFT) and min-Hash techniques for retrieval.

## 3. Research Framework

Proposed algorithm is fast and effective online image search algorithm based on one query image that capture user's search intention. This approach requires user to provide only one query image and images from a pool retrieved by text-based search are re-ranked based on their visual and textual similarities to the query image . The proposed method uses visual vocabulary of vector quantized local feature descriptor (SIFT) to find similarity measures to evaluate near duplicate image detection. Using this duplicate detection technique with existing Internet search system improves precision of top ranked images as result demonstrates.

### 3.1 Adaptive Similarity

A set of visual feature is designed to describe different aspects of image. The Adaptive Similarity is introduced with idea that a user always has precise intention when submitting a query image. For instance, a picture with a big face in the middle most is submitted by the user, most likely user requires images with similar faces and using face-related features is more appropriate. If scenery image is submitted, using scene related feature is more appropriate. The query image is first categorized into one of the predefined adaptive weight categories. There are five types of categories as general object, object with simple background, people, portrait and scene. 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.

### 3.2 Keyword Expansion

User entered query keywords tend to be short and some significant words may be missed because of users' lack of knowledge on the textual description of target images. To capture users' search intention query keywords are expanded, inferred from the visual content of query images, which are not considered in traditional approaches. A word w is recommended as an expansion of the query if a cluster of images are visually similar to the query image and all contain the same word w. With help of visual content and textual description expanded keywords better capture user intention

### 3.3 Image Pool Expansion

Reranking images in the pool retrieved by text-based search is not very effective because image pool accommodates images with a large variety of semantic meanings and the number of images related to the query image is small. Thus, query by keywords should be more accurate which narrow the intention and retrieve more relevant images. Keyword expansions using both visual and textual information retrieve relevant images, which are added into the text query and enlarge the image pool automatically.

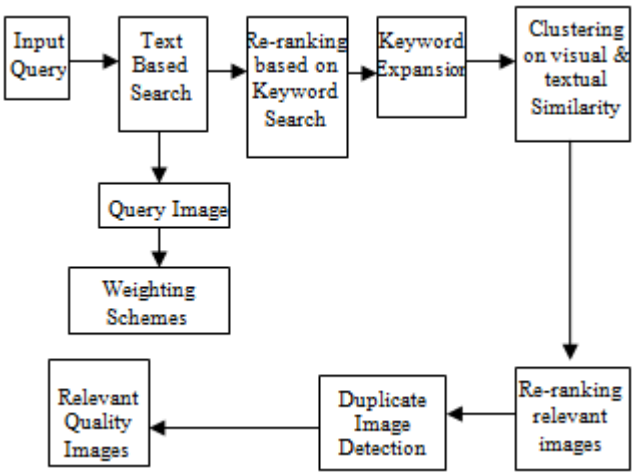


**Figure 3:** Proposed System Architecture.

### 3.4 Visual Query Expansion

To capture search intention one query image is not enough. In keyword expansion step, 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 reranking, because they are more specific and robust to the query image.

### 3.5 Duplicate Image Detection

The similarity measures are applied and evaluated in the context of near duplicate image detection. The proposed method uses a visual vocabulary of vector quantized local feature. For duplicate image detection, 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. Then certain threshold is applied to determine which of the candied images duplicates of query image are.

## 4. Conclusions

This survey provides an overview on the functionality of content based image retrieval system. Most systems are based on low level features such as color and texture, few systems uses shape and spatial location features. Semantic gap can be reduced using various techniques as, object ontology, machine learning methods and relevance feedback. Image search is a process of retrieving and displaying relevant images based on user query. Image search result consist of duplicate images, duplicate can be exact duplicate, global duplicate or near duplicate. The proposed system contains five steps to retrieve relevant images as per user intention. To combine visual feature and to compute visual similarity adaptive to query image an intention specific weight schema is used. To capture user's search intention visual and textual expansions are integrated without additional human feedback. Image pool is enlarged using expanded keywords, to include more relevant images. One short come of the Intent search system is improved using duplicate image detection. To further improve the quality of re-ranked images, this work can be combined with photo quality assessment framework to rerank images by visual quality of the image.

## References

[1] S.K. Chang, S.H. Liu: Picture indexing and abstraction techniques for pictorial databases, IEEE Trans. Pattern Anal. Mach. Intell. 6 (4), 1984.
[2] Yong Rui and Thomas S. Huang, and Shih-Fu Chan: Image Retrieval: Current Techniques, Promising Directions, and Open Issues, Journal of Visual Communication and Image Representation 10, 39–62, 1999.
[3] R.S.Torres, A. X.: Based Image Retrieval: Theory and Applications, RITA, 2006
[4] Ying Liu, Dengsheng Zhang, Guojun Lu Wei-Ying Ma: A survey on content-based image retrieval with high-level semantics, Pattern recognition 40, 2007.
[5] R. Mehrotra, J.E. Gary: Similar-shape retrieval in shape data management, IEEE Comput. 28 (9) (1995) 57–62.
[6] Y. Song, W. Wang, A. Zhang: Automatic annotation and retrieval of images, J. World Wide Web 6 (2) (2003) 209–231.
[7] W. Ren, M. Singh, C. Singh: Image retrieval using spatial context, Ninth International Workshop on Systems, Signals and Image Processing (IWSSIP'02), Manchester, November, 2002.
[8] V. Mezaris, I. Kompatsiaris, M.G. Strintzis: An ontology approach to object-based image retrieval, Proceedings of the ICIP, vol. II, 2003, pp. 511–514.
[9] T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.
[10] E. Chang, S. Tong: SVM active-support vector machine active learning for image retrieval, Proceedings of the ACM International Multimedia Conference, October 2001, pp. 107–118.
[11] Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra: Relevance feedback: a power tool for interactive content-based image retrieval, IEEE Trans. Circuits Video Technol. 8 (5) (1998) 644–655.
[12] X.S. Zhu, T.S. Huang: Relevance feedback in image retrieval: a comprehensive review, Multimedia System 8 (6) (2003) 536–544.
[13] Sagarmay Deb, Yanchun Zhang: An Overview of Content-based Image Retrieval Techniques, 2004 IEEE.
[14] Xin-Jing Wang, Lei Zhang, Ce Liu: A Survey: Duplicate Discovery on 2 Billion Internet Images,
[15] Nadav Ben-Haim, Boris Babenko, Serge Belongie: Improving web-based Image search via content based clustering,
[16] Li Chen, F. W. M. Stentiford: Comparison of near duplicate Image matching,
[17] Ondrej chum, James philbin, Andrew zisserman: Near duplicate image detection: min-Hash and tf-idf weighting, BMVC 2008.
[18] J. Cui, F. Wen, and X. Tang: Real Time Google and Live Image Search Re-Ranking, Proc. 16th ACM Int'l Conf. Multimedia 2008.
[19] Xiaoou Tang, Ke Liu, Jingyu Cui, Fang Wen and Xiaogang Wang: IntentSearch: Capturing User Intention for One-Click Internet Image Search, IEEE transaction on pattern analysis, July 2012.