Web Image Retrieval Using Visual Dictionary

Umesh K K[1] and Suresha[2]

[1]Department of Information Science and Engineering,
S J College of Engineering, Mysore, India.
umeshkatte@gmail.com
[2]Department of Studies in Computer Science, Manasagangothri,
Mysore University, Mysore, India.
sureshabm@yahoo.com

ABSTRACT

In this research, we have proposed semantic based image retrieval system to retrieve set of relevant images for the given query image from the Web. We have used global color space model and Dense SIFT feature extraction technique to generate visual dictionary using proposed quantization algorithm. The images are transformed into set of features. These features are used as inputs in our proposed Quantization algorithm for generating the code word to form visual dictionary. These codewords are used to represent images semantically to form visual labels using Bag-of-Features (BoF). The Histogram intersection method is used to measure the distance between input image and the set of images in the image database to retrieve similar images. The experimental results are evaluated over a collection of 1000 generic Web images to demonstrate the effectiveness of the proposed system.

KEYWORDS

Content-based image retrieval, Dense SIFT feature, Quantization, Bag-of-Features, Bag-of-Words, Similarity measure.

1 INTRODUCTION

Due to rapid changes in digital technologies, in the recent years many people wish to publish their digital information on the Web such as text, image, video, audio etc. Hence, it requires effective indexing and searching tools for Web. Many researchers have been involved in developing the system to retrieve the set of similar Web images, for the given query image. The contents of an image have been used to represent the image semantically. The derived image features are used to retrieve relevant images semantically from the Web.

In the past few years, many researchers have been involved in the area of Content-Based Image Retrieval (CBIR) system to develop techniques to retrieve unannotated images [1]. Today many people use a digital images and video libraries as the main source of visual information. Hence it is an open challenge for the research community to develop cost effective technologies for retrieving, managing and browsing the images in the Web.

Many CBIR systems have been proposed in recent years. The first well known CBIR system is Query by Image Content (QBIC) [2] developed at the IBM Almaden Research Center. Other systems include MIT’s Photobook [3] and its recent version, FourEyes [4], the search engine family of VisualSEEk, WebSEEk, and MetaSEEk [5, 6, 7] (all are developed at Columbia University). The virage [8] is a commercial content-based search engine developed at Virage Technologies. In

DOI : 10.5121/ijwsc.2012.3307

these models, the low level features are extracted from the collected images to retrieve set of relevant images. The results of the system are based on the previously indexed detection results.

The limitations of existing systems are: It is necessary of creating an endless number of object detectors, each requiring ground truth labels, parameter selection, validation, and so on and the similarity-based image retrieval approach requires the user to present one or more images of what he/she is searching using low level features.

Our work is based on the bag-of-features approach. The basic idea of this work is that a set of local image blocks is sampled and a vector of visual descriptors is evaluated on each independently by using Dense Scale Invariant Feature Transform (DSIFT). The resulting distribution of descriptors in descriptor space is then quantified by using proposed vector quantization. As per the survey, there are many clustering/vector quantization algorithms developed to build the visual vocabulary [9,10,11]. In our research work, we proposed novel approach to generate codebook and it reduces the .

According to the literature survey, the bag-of-features were first devised in NeTra image retrieval system [18], which uses codebook to efficiently index color, texture and shape descriptors. The same idea was revisited and implemented in RETIN system [19]. These systems were based on global descriptors. The codebook of local descriptors used in Video Google approach [20]. In the same line, we focused to build visual dictionary of local descriptors, since building a good visual dictionary is the biggest challenge when employing the techniques. The creation of visual dictionary requires the quantization of the description space, which can be obtained by a clustering algorithm. In commonest, the choice found in the literature survey is a combination of dimensionality reduction using Principal Component Analysis (PCA) and a clustering using k-means algorithm with Euclidean distance. This choice has also been criticized [21], since the solution is far from adequate to deal with high dimensional spaces. Also the optimality criterion of PCA does not match the needs of clusterization. Since it chooses the components which maximize the global variance of the data, not the ones which better preserve the local clusters.

The Harris corner detector is very sensitive to changes in image scale, so it does not provide a good basis for matching images of different sizes [22]. Scale Invariant Features Transform (SIFT) approach extended the local features to achieve scale invariance and local image distortions [22]. The novelty of our work is as follows:

1. Dense SIFT (DSIFT) feature description is used to prove that better local feature extraction technique compare with global color model.
2. The critics of basic clusterization technique is considered and proposed novel quantization algorithm for generating good vocabulary dictionary.

The organization of this paper is as follows: In Section 2, the proposed methodology is discussed. The Section 3 describes the experimental results. In Section 3.2, we will conclude our research findings.

2 THE PROPOSED METHODOLOGY

This section presents a brief description of the datasets used in our study, the feature extraction, quantization and distance measures, Bag-of-Features and image representations, and image retrieval using Bag-of-Words. The block diagram of the proposed model is as shown in Figure 1.

2.2 Feature Extraction

The Scale-Invariant Feature Transform (SIFT) extracting keypoints are invariant to rotation, scaling, and translation. These keypoints are used to detect distinctive edges and textures in an image. Each keypoint has 132 dimensions: 128 spatial orientations bins, plus coordinates, scale, and rotation of the keypoints. We will also discuss to extract global image descriptors and represent images as a collection of local properties and calculate from a set of small sub-images called blocks. Image features are extracted from a set of collected images and stored in a database. According to David Lowe [23], SIFT gives a good result to recognize/retrieve images. This approach achieves scale invariance and also less sensitive to local image distortions such as 3D viewpoint change. The feature vector consists of SIFT features computed on a regular grid across the image Dense SIFT (DSIFT). These feature vectors are quantized into visual words to build visual dictionary using proposed quantization algorithm.

DSIFT algorithm makes some new assumptions:

1. The location of each keypoint is not from the gradient feature of the pixel, but from a prede fined location.
2. The scale of each keypoint is same, which is also predesigned.
3. The orientation of each keypoint is always zero.

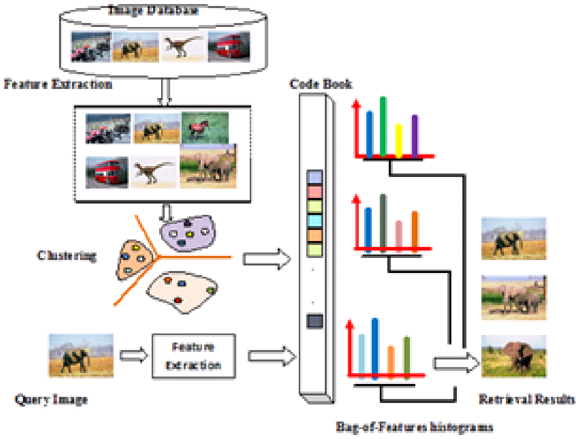

Fig.1. The Block Diagram of Proposed Model.

With these assumptions, DSIFT can acquire more feature in less time than SIFT does.

2.2.1 Dense SIFT

The Dense SIFT feature extraction algorithm is used to extract Web image features. The important parameters of extraction of features from the Web images are the size of the blocks (N) and their spacing (M) (which controls the degree of overlap). In this connection, the RGB component of Web images are converted into to HSV color space and split the image into blocks formed by an N ×N grid. We then compute the mean and variance of each block's color band. We use 4×4 grid in our test, yielding a 16×3×2 = 96-dimensional feature space. The block size tested are N = 5, 7 and 11. The blocks are spaced by M pixels on a regular grid. The blocks do not overlap when M = N, do overlap when M=3 (for N = 5, 7) and M = 7 (for N = 11). The proposed image retrieval system is compared with baseline global color model which is discussed below. This is used to gauge the precision and recall of retrieval tasks.

2.2.2 Global Color Model

The Global Color Model is used to compute the global HSV color histogram for each training Web image. The color values are represented by a histogram with 36 bins for H, 32 bins for S, and 16 bins for V, giving 84-dimensional vector for each image. A query image is tested using histogram intersection and measure the precision and recall separately. In the formulation of histogram, we compute the frequency of each visual word, where each image is represented as a collection of visual words, provided from visual vocabulary. This visual vocabulary is obtained by vector quantizing descriptors computed from the image database. The DSIF descriptor and HSV color space are used as input in proposed novel quantization algorithm separately to generate the vocabulary discussed in Section 2.3.

2.3 Quantization and Distance Measures

In this method, initially all the training vectors are treated as Cluster centers. The distance between a training vector Xi and all the other training vectors Yj where i ≠ j is computed as

$$D = \sum_{j=1}^{k} (X_j - Y_{ij})^2 \text{ where } X \neq Y_i \text{ and } 1 \leq i \leq N \qquad (1)$$

The minimum of all the distances is identified and the Cluster Density of the corresponding training vector is incremented by 1. For example, if the distance between the 1st and the 10000th vectors is the minimum, then the cluster density of the 10000th vector is incremented by 1. Similarly the distance between the second training vector and all the other vectors are computed and so on. Whichever distance is minimum the cluster density of the corresponding vector is incremented by one. The above steps are repeated for all the training vectors one by one. Hence if 250 training vectors are closer to any training vector i, then the cluster density of ith vector is 250. The training vectors are sorted in the descending order based on their cluster densities. Finally the M training vectors with the top density values are selected as the seeds for the initial codebook that is to be used as the input for the Web image representation in vector space model.

Steps to generate the codebook

```
Step1:  Input the image
Step2:  Split the image into small blocks of size 4 x 4.
Step3:  Generate the training set of size N.
Step4:  Initialize the Density array of size N with zero.
Step5:  for i=1 to N
Step6:          for j=1 to N
Step7:          if( i≠ j)
Step8:  compute the distance between the vectors i and j using the Eq.(1).
Step9:    End j
Step10. Find the minimum of all the N-1 distances computed and increment the
        corresponding density by 1.
Step11.         End i.
```

2.4 Image Representations with a bag of visual words

The concept of Bag of visual words method is inspired by Bag-of-Words (BoW) [22], which has been used in text mining. In the BoW model, each stop word is treated as independent feature and achieved outstanding performance in the text retrieval system. The same idea is used in our algorithm to represent Web images semantically. In the Bag of visual model, each image is represented as a set of orderless visual words. Recently few researchers have demonstrated its effectiveness in semantic based image retrieval [23]. The local descriptors (features) are much more precise and discriminating than global descriptors. When looking for a specific image in a image database or target object within the image, this discrimination is very essential in retrieval system, but when looking for complex categories in a large image database, it becomes difficult. To minimize this problem, there is scope of improvement to propose a possible solution is the technique of visual dictionaries. It is made up of set of bag of visual words constructed using clustering approaches.

In the visual dictionary label representation, each region of an image becomes a visual “word” of the dictionary. The idea is that different regions of description space will become associated with different semantic concepts. For example, sky, clouds, land, rocks, vegetation etc having its own semantics. In CBIR, the low level representation considers the descriptor space and split into multiple regions. It employs unsupervised learning techniques like clustering. The limitation of CBIR is that, it describes using low level features. These features are not able to fill the gap between high-level representation and low-level representation. Hence, we proposed Bag-of-Words model derived from text retrieval system to construct a visual dictionary. Based on the visual vocabulary (codebook), each Web image is represented of the corresponding codeword semantically to retrieve similar images from the Web. Besides this compact representation, a sparse but uniform representation is also utilized. In this representation, an image is a vector with each dimension corresponding to a codeword.

2.5 Image Retrieval Using Bag-of-Words

We use histogram intersection method to retrieve similar images from the image database. It is used to compute the similarity between two spatial histograms of given images A and B. The histogram intersection is defined in Equation 1.

$$d(A,B) = 1 - \sum_{i=1}^{m} \sum_{j=1}^{n} \min(a_i, b_j) \qquad (1)$$

where a_i and b_j represent the frequencies of visual words of each block (the whole image is divided into 4 big blocks) of image A and B respectively. The frequency of each visual word is recorded in a histogram for ‘m’ blocks of a spatial tiling. For a given query image Q, the distance between Q and each image in the Web image database will be calculated. Consequently, a set of images in Web with small similarity distance is selected and ranked from the most to the least similar ones.

The quantitative measures of retrieval system are the average precision and recall [12, 13, 23]. We have used the same in our experiment, which is defined in Equations 2 and 3.

$$P(i) = \frac{1}{M} \sum_{j=1}^{M} r(i, j) \qquad (2)$$

$$R(i) = \frac{1}{N} \sum_{j=1}^{N} r(i, j) \qquad (3)$$

where $$r(i, j) = \begin{cases} 1, & if \ id\ (i) = id\ (j) \\ 0, & otherwise \end{cases}$$

where *P(i)* is the precision of query image *i*, *R(i)* is the recall of query image *i*, *id (i)* and *id (j)* are the category ID of image *i* and *j*, respectively, which are in the range of 1 to 10 categories. *M* is the original size of the category that image *i* is from. This value is the percentage of images belonging to the category (relevant) of image *i* in the first *M* retrieved images. The recall is the percentage of images belonging to the category (relevant) of image *i* in the *N* relevant images.

For example, if the query image is a dinosaur (Figure 1), if 89 of the first 100 retrieved images are belonging to the category of dinosaurs, the retrieval precision is 0.89%. Similarly, if the query image is an elephant (Figure 1), if 59 of the first 100 retrieved images are belonging to the category of elephant, the retrieval precision is 0.59% in semantic representation. In global color model, the retrieval precision of dinosaurs is 0.85% and elephant is 0.58%. The average precision of global color model is lesser than DSIFT features semantic representation. Also we observed that in Table 1, the average precision of global color model is 0.49% and the average precision of DSIFT model is 0.62%. Based on the results as mentioned in the Table 1, DSIFT gives better semantic representation than global color model.

Table 2 Comparisons of average retrieval precision (ARP) and precision obtained by proposed methods to other existing image retrieval systems.

Class	Global Color Model		Dense SIFT Color Model	
	Precision	Recall	Precision	Recall
Africa	.44	.70	.58	.78
Beaches	.38	.69	.52	.76
Buildings	.39	.58	.57	.69
Bus	.61	.64	.77	.89
Dinosaur	.84	.85	.89	.90
Elephant	.54	.58	.59	.66
Flower	.67	.67	.78	.83
Horses	78	.73	.89	.89
Mountains	.49	.53	.59	.62
Food	.49	.61	.62	.71

3. EXPERIMENTAL RESULTS

We tested a benchmark Corel image dataset, which comprises of 1000 images from 10 categories [16, 17]. Most of the images are in color photographs in JPEG formatwith the size of 256×384 or 384×256 resolutions. We analyzed the images belonging to different categories and not to a specific domain. This dataset has been used in SIMPLIcity [14] and ALIPR [15] projects to demonstrate the results. We also used WBIIS image dataset, which comprises of 10,000 images from 100 categories [16, 17]. Most of the images are 24-bit true color JPEG format images with the size of 85×128 or 128×85 resolutions. The Corel image datasets are used in our experiments to measure the effectiveness of the retrieval system.

The SIFT features are given as input to proposed Quantization algorithm to visualize and interpret large high-dimensional data sets. First, we investigate the codebook size on retrieval performance of the system. We conducted many experiments to fix the size of the codebook from 50, 100, 150, 200, 250. The 200 size codebook is the best for Corel image dataset and 1000 size codebook is the best for WBIIS low resolution Web crawled image dataset. The detailed results of using SIFT features for each of the 10 categories are shown in Table 1. For the SIFT, we tested block of each region of 16×16 and 8×8 with different size of codebook, the best results are obtained with patch size of codebook. The best results are obtained with patch size of 8×8 and N = 200, which are listed in Table 1. The comparative results of average precision of retrieval results in global Color/DSIFT model for 10 categories are shown Figure 2.

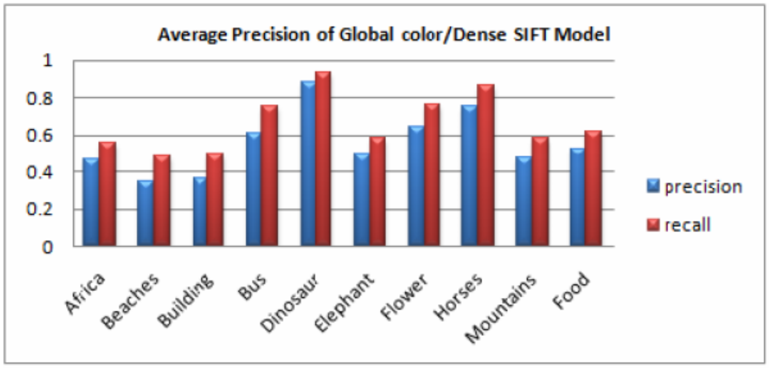

Fig.2. Comparison of average precision of retrieval results in global color/Dense SIFT model for 10 classes

3.1 Performance

The proposed system has been used for feature extraction, generating visual dictionary, image representation and similarity measures. The algorithm has been implemented on an IntelR coreTM i3 processor with 2.20 GHz using Windows operating system. To compute feature vectors for 1000 color images of size 256×384 or 384×256, takes approximately 30 minutes. On average, it takes 1.8 seconds to compute the feature vector of an image. To generate code words, it takes about 5 minutes. The matching speed is quite fast in this architecture. If the query image is in the database, it takes about 0.1 seconds of CPU time on an average to intersect query and indexed images, then sort the distance of all the images to be demonstrated at the user side. The new image retrieval system is compared with baseline global color model. These are included in order to gauge the precision of retrieval tasks. By using the visual dictionary techniques, the applications have been greatly broadened and encompass category retrieval in image databases.

3.2. Conclusions

We have proposed semantic based image representation to retrieve images from the Web image database. The proposed quantization algorithm is an iterative procedure used to generate code words for every image to represent cluster similar features efficiently. The results of our experiments show that the proposed technique is able to retrieve similar kind of image effectively from the Web image database. In Web to increase better retrieval performance, we have to do an extensive study of different feature representations to reduce quantization error and topological error rate to evaluate the complexity of the output space. We observed that, large numbers of unclassified images are available on the internet, which are the major challenges of collection of Web images. We conducted an Experiment on a Corel Web image database to demonstrate the efficiency and effectiveness of the proposed framework.

REFERENCES

[1] Rui, Y., Huang, T.S., Chang, S.-F.: Image Retrieval: Current Techniques, Promising Directions, and Open Issues. Journal of Visual Communication and Image Representation. 10 (1), 39-62 (1999).

[2] Flickner, M., Sawhney, H., Niblck, W., et al.: Query By Image and Video Content: the QBIC system. IEEE Computer September, 23-31 (1995).

[3] Pentland, A., Picard, R.W., Sclarof, S.: Photobook: Tools for Content-Based Manipulation of Image Databases. In: Storage and Retrieval for Image and Video Databases II. In: SPIE Proceedings Series, Vol. 2185, (1994).

[4] Minka, T.P.: An Image Database Browser that Learns from User Interaction. Master's thesis, M.I.T., Cambridge, MA (1996).

[5] Michael J. Swain., Charles Frankel., and Vassilis Athitsos.:WebSeer: An Image Search Engine for the World Wide Web, Technical Report 96-14, (1997).

[6] J. R. Smith,: Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis. PhD thesis, Graduate School of Arts and Sciences, Columbia University (1997).

[7] S. Sclaroff., L. Taycher., and M. La Cascia.: Imagerover: A Content-Based Image Browser for the World Wide Web. In Proceedings IEEE Workshop on Content-based Access of Image and Video Libraries (1997).

[8] Bach, J.R., Fuller, C., Gupta, A., et al.: The Virage Image Search Engine: an open framework for image management. In: Sethi, I.K., Jain, R.J. (Eds. , Storage and Retrieval for Image and Video Databases IV. In: SPIE Proceedings Series, Vol. 2670. San Jose, CA, USA (1996).

[9] Koikkalainen, P.: Progress with the Tree-Structured Self Organizing Map. In: Cohn, A.G. (Ed.), 11th European Conference on Artificial Intelligence. European Committee for Artificial Intelligence (ECCAI). Wiley, New York (1994).

[10] Koikkalainen, P., Oja, E.: Self-Organizing Hierarchical Feature Maps. In: Proceedings of 1990 International Joint Conference on Neural Networks, Vol. II. IEEE, INNS, San Diego, CA (1990).

[11] T.Kohonen, Self-Organizing Maps, Springer-Verlag, New York (1997)

[12] Salton, G., McGill,M.J.: Introduction to Modern Information Retrieval. In: Computer Science Se-ries. Mc- Graw-Hill, New York (1983).

[13] Li, J.,Wang, J. Z. andWiederhold, G.: Integrated Region Matching for Image Retrieval. ACM Multimedia, p. 147-156 (2000).

[14] James Z. Wang, Jia Li and Gio Wiederhold.: SIMPLIcity: Semantics-Sensitive Integrated Match-ing for Picture Libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 9, pp. 947-963 (2001).

[15] http://alipr.com/

[16] R. Datta, J. Li, and J. Z.Wang.: Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition. Proc. IEEE ICIP, Special Session on Image Aesthetics, Mood and Emotion, San Diego, CA (2008).

[17] http://wang.ist.psu.edu/jwang/test1.jar.

[18] W-Y. Ma, B.Manjunath.: NeTra: A Toolbox for Navigating Large Databases. Multimedia systems, Vol. 7, n.3, pp. 184-198 (1999).

[19] J.Fournier, M.Cord, S. Philipp-Foliguet.: RETIN: A Content-Based Image Indexing and Retrieval System. Journal of Pattern Analysis and Applications, Vol. 4(2/3), pp. 153-173 (2001).

[20] J.Sivic , A.Zisserman.: Video Google: A Text Retrieval Approach to Object Matching in Video. In Proc. 9th IEEE International Conference on Computer Vision, pp. 1470-1477 (2003).

[21] B. Triggs, F.Jurie.: Creating Efficient Codebooks for Visual Recognition. In IEEE Int. Conf. on Computer Vision, vol. 1, pp. 604-610 (2005).

[22] Q.Tian, S.Zhang.: Descriptive VisualWords and Visual Phrases for Image Applications. ACM Multimedia, pp. 19-24 (2009).

[23] David G.Lowe.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60, pp.91-110, 2(2004).