*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. I (July – Aug. 2015), PP 71-76*
*www.iosrjournals.org*

# A Review onImage Mining Techniques and its application on a software BOND

[1]Shubham Krishna ,
*1Student, Bachelor of Technology,Department of Computer Science,2SVKM's NMIMS, Mukesh Patel School of technology Management and Engineering,*

[2]Vatsal Gosalia
*[2]Student, Bachelor of Technology, Department of Computer Science,*
*2SVKM's NMIMS, Mukesh Patel School of technology Management and Engineering,*
*Shri Bhakti Vedant Swami Marg, Vile ParleMumbai, India.*

***Abstract:*** *Image processing is one of the most researched areas in computer science and it finds numerous applications in various fields like, medical research and diagnosis, geological research, crime investigation, and so on. In most of the institutions, authentication of members is carried out by facial recognition, fingerprint or retina scan, etc. Recently image mining has gained interest of innumerable researchers because of its widespread applications and so little development in the field. This paper focusses on studying about some of the important techniques researched and developed successfully to support image mining with considerable accuracy. In later sections, a comparative study of the same is carried out, which tries to identify the best method of all.*

***Keywords:*** *Feature Extraction, Image Mining, Image retrieval, Object Recognition, Image Clustering*

## I. Introduction

Image mining refers to maintaining and extracting of images, generally huge set of images. Digitization of data has opened up the gates for storage and access of huge amount of data, image being the better part of it. However, the complete automation of image based searches has yet not been achieved. By the help of some vectors, such as color, texture etc., the library of images can be very efficiently used. Image searching deals with extraction of features from the image.

Feature extraction is the most crucial step in any image mining algorithm. Feature Extraction is the process of extracting one or more features from the input image and comparing them with the features of database images, in order to find a perfect or approximatematch. The basic approach is to sort images according to some important features like: color, texture, shape, temporal details, edge, shadows and other temporal details and all of this should be carried out with minimal human interaction.

The following sections will cover some techniques aiming at efficient mining of images from a huge database:

## I. Algorithms

### 1. Feature Extraction Based Technique:

#### 1.1 Color feature extraction:

Some of the techniques tried were – Average color in Gray scale, Average color in RGB format [GW92] and Average color in YCBCR (Y is the luminance and CB, CR are the chrominance components) [GW92].[1]

There were two general methods involved in the color feature extraction which were precision and recall.Precision is that part of the feature directory which is retrieved from the training image and that is relevant amongst all the provided features. However, the YCBCR method was found to be a bit more efficient.

There was a general formula introduced for the same:

$$\text{Average color} = \frac{\sum(\text{intensity of all the pixels in the current block})}{\text{total number of pixels in the current block}} [1]$$

Now, the output which will be generated from this operation will be a new matrix which has 0's at the places where the color match was found false and 1 where the match was found true.

Thus, the test for color match was made.

#### 1.2 Texture feature extraction:

For texture extraction, three different types of images were chosen: image with only egeria, image with egeria and water and image with egeria and land to cover all scopes of images. Histograms ( histograms without bins, with bins [JV96], Normalized Histogram with bins[JV96] and Discrete Cosine Transform [GW92] of these different images were obtained and the difference between the peaks of these histograms were taken. [1] Wherever the distance was found to be relatively smaller and matches the feature matching criteria (threshold) is extracted. The formula used is:

$$\text{Similarity measure (d)} = \frac{\sqrt{\sum(\text{Means for bins for template} - \text{Means for bins of block})}}{}[1]$$

Here, similarity measure is the relative distance between the histogram peaks of the query block and that of the template.

### 1.3 Edge feature extraction:

Edge extraction is amongst the most important steps of image comparison. Edges are high frequency regions. This is because introduction of an edge means there is some admirable intensity change in the image, which means a large change in the frequency of the image. Edge features are particularly important for some of the darker images.[1] since edge feature extraction is not powerful enough, it has to be proceeded with another method like color or feature extraction.

Hence, all these operations were performed jointly to get the end result. The grouping of these actions together calls for the most accurate image pattern recognition and also accounts for the ease to store and search images.

## II. Block Truncation Algorithm

Before moving ahead with the working of this algorithm, we first need to learn about some prerequisite terms. Those terms are color moments and clustering. Now there are many clustering techniques, so we shall have a look only at the most widely used one, i.e. the K-means clustering. This would help get a better understanding of the algorithm.

### 2.1 Color Moments

If the colors present in an image can be viewed as a probability distribution, then the moments of that calculated probability distribution can be used as a feature in image grouping. In simple words, for a given image, the moments of a probability distribution can act like a feature in itself and can thus be used in identifying the image based on color.

The color feature extraction can be done very efficiently using 3 major moments: mean, standard deviation, and skewness.

$$E_i = \frac{1}{m.n}\sum_{j=1}^{m.n} P_{ij} \quad \ldots (1)$$

$$\sigma_i = \left[\frac{1}{m.n}\sum_{j=1}^{m.n}(P_{ij} - E_i)^2\right]^{1/2} \quad \ldots (2)$$

$$\alpha_i = \left[\frac{1}{m.n}\sum_{j=1}^{m.n}(P_{ij} - E_i)^3\right]^{1/3} \quad \ldots (3)$$

Figure: 1 [2]

where, $E_i$, $\sigma$and $\alpha_i$ represent the mean, standard deviation and skewness of each color channel (r,g,b). $P_{ij}$ is the value of each color channel at $j^{th}$ pixel. The product m.n gives the total number of pixels per image.

### 2.2 K-means Clustering

Being one of the simplest of the algorithms, the k means clustering algorithm involves each pixel to be put into a single category of its own. The user has to choose the number of clusters (k). Then, we find the centroid co-ordinate, which actually represents the mean of the pixels. Then, iteratively, the pixels are subtracted from this centroid, or the distance of a pixel from this centroid was taken and the pixel with the least distance was grouped with all such pixels having similar distance from the centroid. This way, clusters or groups were developed.

**2.3 Algorithm**

The block truncation algorithm follows these simple steps:

1. Divide the given image into 3 components viz, Red, Blue and Green
2. Compute the average of each component
3. Compute the

i. $R_H$,$R_L$
ii. $G_H$,$G_L$
iii. $B_H$,$B_L$

Where, the subscript H implies High, signifying that the average of all the redcomponents from the query image is greater than the component average and L implies low, which is obtained by taking only that component which are below that component average.

**2. Feature cascading approach towards object recognition**

This is quite a unique approach in itself, since it doesn't involve only one iteration of feature extraction but cascades several of them. The work in this method involves 3 important steps. The first one involves preparation of a new type of representation of an image known as "Integral Image". This term will be explained in detail later; in fact, every step will be. This integral image is then passed through a machine learning algorithm, based on AdaBoost, that performs the task of extracting all the features from the integral image, but rejecting most of the unimportant ones, thus leaving only the critical features to be passed on to step-3. [3]The structure and function of the processing in step-3 is what makes this approach unique. This step involves cascading of incrementally complex classifiers in a series connection, thus discarding more and more unwanted features at each block in the cascade. A detailed and illustrated example of each of the steps will clear the fog.
Before proceeding with the detailed explanation of all the 3 steps, we first need to take a glance at rectangle features of an image.

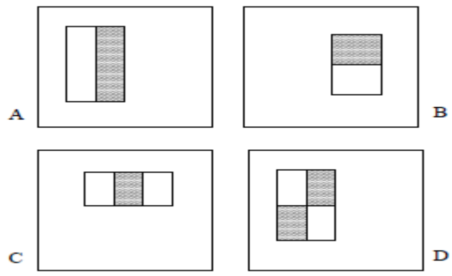


**Figure – 2:** Example of rectangle features [3]

Blocks A and B in the above figure show two-rectangle features, block C shows a three-rectangle features and the block D shows a four-rectangle feature. The sum of the pixel values in the white portion is subtracted from those in the black portion. These rectangle features prove to be very useful in detection of objects that exhibit a heterogeneous color and texture pattern, such as a human face.
So now, time to proceed with the 3 main algorithmic structures.

**Step-1: Preparation of the integral image.**

Instead of working directly with image pixel intensities, the algorithm focuses on working with a special type of image, known as integral image. Okay, enough suspense now. Integral Image, or rather I would say, integral image pixel value at a given position (x,y) is the contains the sum of pixel values that are above and below that position in the original image. We always tend to better understand equations, so here it is:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x', y'),$$

[3]

Using this integral image, we can realize a very rapid computation of rectangular features. [3]

**Step – 2:Feature Extraction and validation.**

Within image sub-window, the total number of Harr-like features is very large, making it utterly important to focus on some small set of significant features, in order to ensure economy, in terms of space and time. Now is the time when the rectangle features described previously, come into play. The value of two-rectangle featuresis the difference between the sum of the pixels within two rectangular regions. The three-rectangle features is the sum of pixel values within two outside rectangles subtracted from the sum of pixels in a center rectangle. And the four-rectangle feature is the difference between sum of pixel values in a diagonal pair of rectangles.

Note that the system proposed in the paper is a variant of a primitive system known as AdaBoost. Now, even though each feature for a given image sub-window can be computed efficiently, computing the complete set proves to be expensive. In order to avoid this, a weak learning algorithm is designed to select the single rectangle feature which best acts as a threshold between interesting and uninteresting features.
A weak classifier $h_j(x)$ will hence contain a feature $f_j$, a threshold $T_j$ and a parity $P_j$ indicating the direction of the inequality sign: [3]

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < p_j\theta_j \\ 0 & \text{otherwise} \end{cases}$$

[3]

Where ‘x’ is a 24x24 pixel sub-window of an image.

**Step – 3: The Attentional Cascade**

This section focuses on preparing an incremental cascade structure of classifiers that goes on discarding useless features, thereby radically boosting the computational efficiency at each stage. The cascade is so formed that simpler classifiers initially reject most of the sub-windows before more complex classifiers are called upon to reduce false positive rates. [3]A positive result i.e. a green signal from the first classifier triggers the evaluation of the second classifier and so on. A negative result at any point results in immediate termination of the process for that feature and thereby, the rejection of the sub-window.

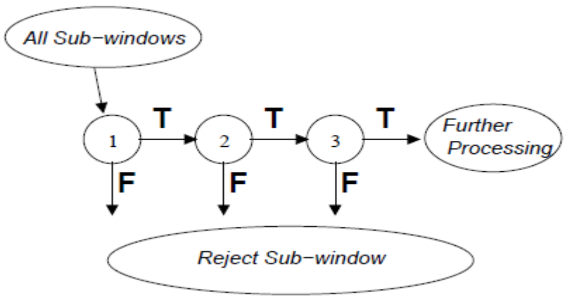


**Figure – 3:** Cascade of classifiers and its working [3]

## III. Inferences

Overall we saw three different methods for image mining technique. All of the three methods were used for different purposes and were having immense impacts in the classification of features of the different images collected. The feature extraction methods use for the detection of the Egeria Densa was quite effective. Through the use of this technique, the efficiency of the precision and recall methods was found to be as below:
Where, Precision is defined as the fraction of the retrieved information, which is relevant.

$$\text{Precision} = \frac{TP}{TP+FP} \text{ [1]}$$

•And Recall is defined as the fraction of the relevant retrieved information versus all relevant information.

$$\text{Recall} = \frac{TP}{TP+FN} \text{ [1]}$$

The results that we get here are either of the four: true positives, true negatives, false positives and false negatives. True positives (TP) are those extracted regions that are correct, True negatives (TN) are the regions that are incorrect and are not retrieved, False positives (FP) are regions that are actually incorrect, but have been extracted, and FN are regions which were supposed to be extracted but were missed. [1]
Table 1 displayed below shows the experiments based on the precision and recall system:

| Image Name | Calculation | | | |
|---|---|---|---|---|
| | Certain | Uncertain | Precision | Recall |
| lvi2 | 406 | 0 | 0.8188 | 0.9010 |
| bb1 | 338 | 191 | 0.9595 | 0.6614 |
| di1 | 301 | 133 | 0.7181 | 0.7461 |
| hr1 | 248 | 342 | 0.2956 | 0.9309 |
| ft1 | 847 | 29 | 0.3778 | 0.9795 |
| lps1 | 227 | 426 | 0.2032 | 0.9056 |
| ls1 | 281 | 306 | 0.3180 | 0.6807 |
| lvi1 | 150 | 505 | 0.2863 | 0.8909 |
| orh1 | 209 | 466 | 0.1901 | 0.7571 |
| qi1 | 235 | 217 | 0.6099 | 0.8471 |
| vc1 | 213 | 76 | 0.8931 | 0.6383 |
| wi1 | 305 | 539 | 0.5393 | 0.8186 |
| ds1_7-02 | 236 | 107 | 0.1847 | 0.7562 |
| 7ms_7-02 | 467 | 94 | 0.1727 | 0.9785 |
| ft1_7-02 | 471 | 164 | 0.5834 | 0.8381 |
| ft2_7-02 | 494 | 82 | 0.4994 | 0.7247 |

**Table 1: Feature extraction using color, texture and edge features [1]**

The block truncation algorithm was another good and simple method to perform image recognition. The result of the experiment performed with a database of 1000 images is enlisted below:

| *Classes* | *Recall* | *Precision* |
|---|---|---|
| African People and villages | 30 | 25.43 |
| Beaches | 47 | 40.17 |
| Buildings | 33 | 23.57 |
| Buses | 37 | 35.24 |
| Dinosaurs | 100 | 92.59 |
| Elephants | 32 | 35.56 |

**Table 2: Recall and precision using Color Moments [2]**

| Classes | Recall | Precision |
|---|---|---|
| African People and villages | 44 | 33.58 |
| Beaches | 42 | 42.42 |
| Buildings | 8 | 7.92 |
| Buses | 52 | 44.83 |
| Dinosaurs | 99 | 97.06 |
| Elephants | 39 | 44.83 |
| Flowers | 80 | 94.12 |
| Horses | 50 | 58.82 |
| Mountains and glaciers | 35 | 43.21 |
| Food | 25 | 22.32 |

**Table 3: Recall and precision using BTC Algorithm [3]**

The table 3 shows the experimental analysis of the block truncation algorithm which was found to be better than the color moments technique.

The third algorithm studied is a very efficient one in terms of time and space complexity, but when implementation is concerned, the cascade structure of the algorithm could become a bottleneck, as that many functions have to be designed to discard features at early stages.

## IV. Conclusion And Future Scope

Image mining is the backbone for databases consisting of images.With the growing importance of data mining and its use, image mining too finds its respect, as many researchers are interested in this domain. Images can be sorted according to their features, namely, color, texture etc. this makes the searches quicker and easier, therefore, providing for a better experiences when dealing with an apt size of image database. Various methods have been implemented which show promising traits and some of them were found to be quite efficient. In all, image mining has secured its future in many technological areas, like pattern recognition, content based image retrieval (CBIR) and many more. There are many search engines like google which used the image mining techniques for the filtered search of the images. Apart from all this, there are some loopholes still existing in this field which need to be taken care of, like loss of data, adequate accuracy and speed of the search. The goal must be towards achieving these objectives in the near future. The future scope is implemented in our upcoming software called BOND which gives the exact geographical location of any digital image, by extracting its background.We intend to help crime branches to gather geographical information about any suspect digital picture. For getting the latitudinal and the longitudinal details about the picture, our software links to the Google's satellite pictures via Google Earth and will also take help from the Indian Metrological Department (IMD)[5] which will provide with the satellite help and compare the query image with the images on the database of the IMD, gathered through the satellites.

## References


[1]. Feature Extraction for Image Mining Patricia G. Foschi Romberg Tiburon Center for Environmental Studies San Francisco State University Deepak Kolippakkam*, Huan Liu and Amit Mandvikar Department of Computer Science & Engineering Arizona State University, Tempe, Arizona, USA

[2]. Color Image Clustering using Block Truncation Algorithm

[3]. Dr. Sanjay Silakari 1, Dr. Mahesh Motwani 2 and Manish Maheshwari 3

[4]. 1 Department of Computer Science and Engineering Rajiv Gandhi Technical University Bhopal, M.P., India

[5]. 2 Department of Computer Science and Engineering Jabalpur Engineering College Jabalpur, M.P., India

[6]. 3 Department of Computer Science & Application Makhanlal C. National University of Journalism and Communication Bhopal, M.P., India

[7]. Rapid Object Detection using a Boosted Cascade of Simple Features

[8]. Paul Viola Michael Jones viola@merl.com mjones@crl.dec.com Mitsubishi Electric Research Labs Compaq CRL 201 Broadway, 8th FL One Cambridge Center Cambridge, MA 02139 Cambridge, MA 02142

[9]. THE IDENTIFICATION OF PILL USING FEATURE EXTRACTION IN IMAGE MINING A. Hema1 and E. Anna Saro2 1Department of Computer Applications, Manonmaniam Sundaranar University, India 2Department of Computer Applications, S.N.R Sons College, India

[10]. http://www.imd.gov.in/