ISSN: 0976-9102(ONLINE)
DOI: 10.21917/ijivp.2015.0143

THE IDENTIFICATION OF PILL USING FEATURE EXTRACTION IN IMAGE MINING

A. Hema[1] and E. Anna Saro[2]

[1]*Department of Computer Applications, Manonmaniam Sundaranar University, India*
E-mail: hemaresearch@yahoo.com

[2]*Department of Computer Applications, S.N.R Sons College, India*
E-mail: saroviji@rediffmail.com

Abstract

With the help of image mining techniques, an automatic pill identification system was investigated in this study for matching the images of the pills based on its several features like imprint, color, size and shape. Image mining is an inter-disciplinary task requiring expertise from various fields such as computer vision, image retrieval, image matching and pattern recognition. Image mining is the method in which the unusual patterns are detected so that both hidden and useful data images can only be stored in large database. It involves two different approaches for image matching. This research presents a drug identification, registration, detection and matching, Text, color and shape extraction of the image with image mining concept to identify the legal and illegal pills with more accuracy. Initially, the preprocessing process is carried out using novel interpolation algorithm. The main aim of this interpolation algorithm is to reduce the artifacts, blurring and jagged edges introduced during up-sampling. Then the registration process is proposed with two modules they are, feature extraction and corner detection. In feature extraction the noisy high frequency edges are discarded and relevant high frequency edges are selected. The corner detection approach detects the high frequency pixels in the intersection points. Through the overall performance gets improved. There is a need of segregate the dataset into groups based on the query image's size, shape, color, text, etc. That process of segregating required information is called as feature extraction. The feature extraction is done using Geometrical Gradient feature transformation. Finally, color and shape feature extraction were performed using color histogram and geometrical gradient vector. Simulation results shows that the proposed techniques provide accurate retrieval results both in terms of time and accuracy when compared to conventional approaches.

Keywords:

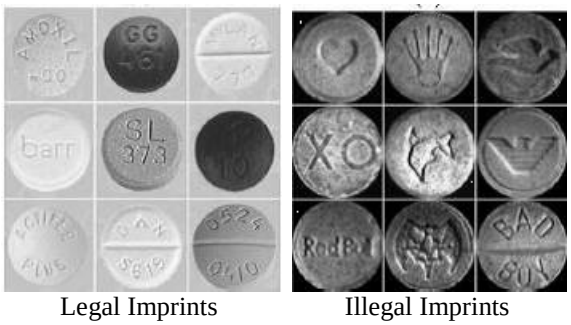
Image Mining, Feature Extraction, Image Matching, Pill Images, Image Retrieval

1. INTRODUCTION

The present scenario of the pharmaceutical industries seems to be growing in a higher pace as there are numerous diseases arising all over the world. The regimen provided by the products of such industries act as a major global health care system.

However, illegal and irrelevant drugs lead to major public hazard. This is attributed to the unawareness due to lack of education and poverty which gives opportunity for emerging of some companies producing low quality drugs to make profit out of their business. The medicines are hence are subjected to authentication by means of the trademarks and characters printed on the surface of the tablet. The legality of the pills/tablets can be identified by the druggist by means of such authentication, yet such preliminary identification alone is not sufficient to confirm whether the pill is legal or illegal. Experimental and instrumentation method of identification are available however, such analysis will not be helpful during criminal investigation or at any crucial stage.

This proposed work aims at identification of pills using image mining methods so as to be suitable at the time of crime scene. Reliable and facile identification method for finding legal and illegal pills can help physicians and patients to decrease uncertainty and thus can also gain patient's confidence towards the healthcare system. The authenticated pharmaceutical companies have enrolled with their identification marks in the database which will provide information to segregate legal pills out of illegal drug pills. Further, law enforcement databases used to have the information recorded if any recent illicit psychoactive pill is initially detected in the marketplace. The legality of each pill can be verified by means of imprint which is a printed mark on the tablet or capsule in terms of digits, characters, letters, symbols or the combination of any of them. Some of the companies that use the symbols on the pill for the sake of advertisement purpose can also be known for its originality. The ministry of regulation also insists and checks for each prescription having unique identification mark in terms of size, color, shape and imprint. After all these validation procedures only can the pill is approved legal by the quality control department in the Ministry of Chemical and fertilizer. The images carrying imprints of the legal and illegal pills are shown below.



Legal Imprints Illegal Imprints

Fig.1. Sample Pill Images

This research work mainly focused on providing efficient drug identification system using image mining techniques for detection, retrieval with matching of the image pills.

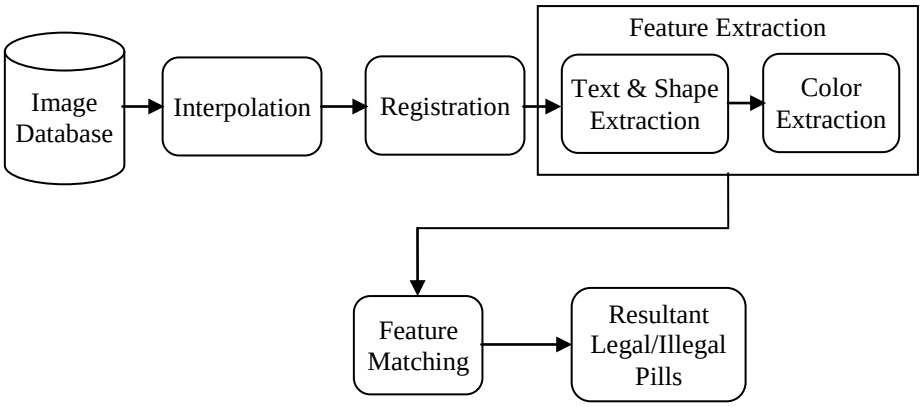


Fig.2. Overall Research Diagram

2. LITERATURE REVIEW

The errors that occur during drug prescription, dispensing or administration is collectively termed as medication errors. The extraction of textual blocks from gray scale document images can be done by the method using edge information [1, 13]. A simple method based on edge detectors such as the Sobel operator for document image filtering and text extraction has also been studied [9]. The Harris corner detector [10] was employed for overcoming the object recognition problem by identifying the interest points followed by creating a local image descriptor a teach interest point from an orientation-invariant vector of derivative-of-Gaussian image measurements. In recent times, the image retrieval from a large image database based on color projections and different statistical approaches are applied [4]. Such approaches as termed as query-by-example as it extracts the images depends on the user request. It is the representative of a visual query to the system used for comparison of some similarity metrics with query and target images [5]. The most common approach to compare the color content of a query image with that of database images is done by using color histograms. It is a familiar fact that a series of pixel values are visible to corresponding color represents the images and this methodology is working on such fact.

3. PROPOSED METHODOLOGY

The overall research work is as follows

3.1 PREPROCESSING

The foremost step before feature extraction is called as preprocessing which makes use of Modified decision based Unsymmetric Trimmed Median Filter (MDBTUMF) to avoid noisy features from the images and storing the remaining features.

3.1.1 Interpolation:

The decomposition of the images into high and low frequencies is known as scaling so that the low frequency pixels which are considered as known signal can be separated from the high frequency unknown signal with the help of decimated wavelet. Interpolation is the process that is performed by scaling the image in terms of reducing the blur, error and jagged edges. The interpolation process is carried out using bicubic interpolation with undecimated wavelet to make the size of the query image similar to database image. After interpolation, overall the size of the query image become similar and hence can be stored as a separate interpolated database for further processing. Later, the registration process will be done to handle the noise and to extract the appropriate features. These features are matched with the characteristics of the original image with respect to color, shape and imprints and the output is investigated to determine whether the input pill image is legal or illicit.

The algorithm starts with the detection of impulse noise. The main aim of introducing this interpolation algorithm is to reduce the artifacts, blurring and jagged edges introduced during up-sampling. The enhanced interpolation algorithm initially uses a region separation procedure that separates the edge and detailed region of the image. After separating the edge and detailed regions of the image, to interpolate the edges, the orientation angle, OA is determined and is quantized with respect to its closest neighbouring pixels. Next, the quantized orientation angle is calculated for each of the original edge pixels. If this angle has the support of more than 60% pixel support, then this value is taken as the dominant edge angle of that block for coordinate rotation. In order to adjust the weights of the reference pixels to smooth the edge and prevent the jagging artifacts, after determining the dominant edge orientation of the sliding block, the original coordinates of the sliding block are rotated to the new coordinates.

It is evident that after rotation the i' axis is parallel to edge and j' axis is normal to the edge. To enhance the edges, the distance between the pixels (supplementary and surrounding) in the i' axis is shrunk and j' axis is stretched for the bicubic kernel function. This makes the weights of the reference pixels is increased along the edge direction and weakens the pixels normal to the edge. Thus the 2-dimensional bicubic interpolation kernel to interpolate a supplementary pixel at coordinate in the image is calculated.

Hence the edge region is interpolated using bicubic algorithm. An Undecimated wavelet based interpolation algorithm is used for the detailed region. Undecimated wavelet (UW) based interpolation is used to enhance the detailed section of the cropped tablet image. The interpolation steps involved during this process is shown in Fig.3. The UW is used to take advantage of the fact that with UW the approximation

coefficients and detail coefficients at each level are the same length as the original signal and hence can provide a denser approximation to the continuous wavelet transform than the approximation provided by the orthonormal discrete wavelet transform (DWT). A quadtree weight function is used to modify the estimated detail wavelet coefficients to exploit the inter correlation between the UW subbands. If X is the input detailed region of the tablet image, the output interpolated image, X', is obtained by applying low pass filtering followed by decimation to give signal A. The original high-resolution image X filters with the high pass filter H to obtain the detail signals D. During enhancement, A is considered as the known signal and D as unknown, which has to be estimated. Finally, after estimation an inverse UW is performed to obtain X'.

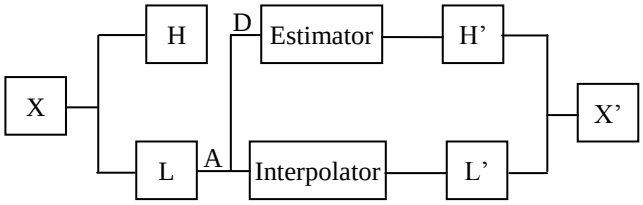


Fig.3. Detailed Region Interpolation

3.2 REGISTRATION

The process of finding the correspondence point between two same scene or object images is known as image registration.

This work proposes a novel method called by coupling the concept of image mining which is an integral part of data mining. The novel approach called Image Registration is proposed here to handle the noise and to extract the appropriate features. Also, the de-noising activity is assisted to discard noisy high frequency edges and to retrieve the valuable high frequency edges. Moreover, three distinct methods like Surf, Haar wavelet and Harris corner are proposed to enhance the performance and to increase the stability.

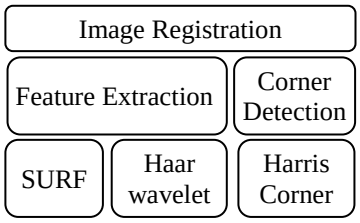


Fig.4. Framework of the Image Registration module

In this registration phase where the noisy features of high frequency edges are discarded and relevant features of high frequency edges are selected. Image registration is defined as the process of transforming different sets of data (photograph images, sensors, times, depths, etc) into one coordinate system. Image registration is the process of finding the correspondence point between two same scene or object images. There are three different concepts associated in finding the correspondences like interest points (detectors), neighbourhood points and descriptor vectors. A wide variety of detectors and descriptors were proposed in the related works but all satisfies in the performance factor, hence the distinctive and fast detector is called SURF (Speeded up Robust Features) is adopted in this research.

Herbet Bay et al (2006) presented and developed a local feature detection method called SURF (Speeded Up Robust Features). For detecting high frequency pixel values in rows and columns of the image, Haar wavelet transform is used which is denoted as shown below:

$$\Psi(t)=\begin{cases}1 & 0\le t<\frac{1}{2}\\ -1 & \frac{1}{2}\le t<1\\ 0 & otherwise\end{cases} \quad (1)$$

The Hessian Matrix which is suitable for fast detection of integral images is employed in SURF to detect the interest point and convert the resultant pixels in the form of vector values. The Hessian matrix $H(x,\sigma)$ in x at scale σ is represented as,

$$H(x,\sigma)=\begin{bmatrix}L_{xx}(x,\sigma) & L_{xy}(x,\sigma)\\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma)\end{bmatrix} \quad (2)$$

where, $L_{xx}(x, \sigma)$ is the complication of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2}g(\sigma)$ in the image I at the point x and similarly are $L_{xy}(x, \sigma)$ and $L_{yy}(x, \sigma)$.

Harris corner detection is based auto-correlation function for measuring the local changes of the input with the features or pixels that are shifted by a small amount in different directions. It is thus defined as given in Eq.(6) where w represents the window size and $I(x,y)$ represents the intensity value. The measured hanges in the intensity for the shift $[u,v]$ is given by,

$$E(u,v)=\sum_{x,y} w(x,y)\left[I(x+u,y+v)-I(x,y)\right]^2 \quad (3)$$

3.3 FEATURE EXTRACTION AND MATCHING

The process of reducing the large data by means of extracting the shape, color and text separately is known as the Feature extraction. The three features that are to be extracted from the interpolated images three features are Text or Imprint, Shape and Color. The use of geometrical Gradient feature transformation algorithm is well described in following steps:

1) Get query image as input
2) Apply gradient method to that image
3) Detect edge pixels by scanning each and every row and column of image
4) Locate center value of pill image
5) Calculate Euclidean distance between center pixel and each pixel in image
6) Obtain the minimum distance from above step
7) Repeat the same process for database image

3.3.1 Text and Shape Feature Extraction:

Geometrical Gradient feature transformation algorithm can be used for extracting the feature Text or Imprint from the query pill. During text matching, the query pill image is considered as an input image. The geometrical gradient feature transform operates by locating he centre pixel position in the foreground of the text surface by considering only the maximum edge values thereby to match the text on pill image with that of the database image. If the edge of the image points in a variety of directions,

canny algorithm can be used four filters to identify horizontal, vertical and diagonal edges of the image. The edge detection operator works by recurring a value for the primary derivative in the horizontal direction (G_x) and the vertical direction (G_y).

The formula for Geometrical Gradient feature transformation is,

Calculating the gradient values- the edges should be marked in the image where the gradient of the image is maximum.

$$G_x = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \text{ and } G_y = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

Gradient magnitude and the angles for the image are calculated which is given by,

$$M[i,j] = \sqrt{P[i,j]^2 + Q[i,j]^2} \text{ and } \theta[i,j] = \tan^{-1}(Q[i,j], P[i,j])$$

$$G = \sqrt{G_x^2 + G_y^2} \quad (4)$$

G_x ,G_y – Horizontal, Vertical direction of edge operator in both row and column wise. The edge pixel is detected and extracted from the text separately by means of scanning. The features of query image are then compared with database image by taking the center pixel for each letter and calculating Euclidean distance between center pixel and remaining pixel value. A method used to find "ordinary" distance between two points is called Euclidean distance and can be measured with a ruler using the Pythagorean formula given below.

Centroid:

$$C = (C_X, C_Y)$$

$$C_X = \frac{X_1 + X_2 + X_3 + \ldots + X_K}{K} \quad (5)$$

$$C_Y = \frac{Y_1 + Y_2 + Y_3 + \ldots + Y_K}{K} \quad (6)$$

where, X and Y are the X co-ordinate and Y co-ordinate positions of all points forming the shape (or along the perimeter of the shape)

$$D = \sqrt{\left(\left(C_X - P_X^I\right)^2 + \left(C_Y - P_Y^I\right)^2 \right)} \quad (7)$$

where, P_X and P_Y is the position of every feature and I is the total number of Features.

The distance between those arranged in vector form is calculated using above formula. The same process is repeated for database image and the difference in distance between these two images is determined. For those texts having different scale and rotation, this method is suitable.

The query image resulting in minimum difference is finally considered as the matching image with that of the database. For extraction of other features such as shape, the same process is repeated. The resultant value in the vector form is found by considering minimum edges instead of maximum for the matched pill.

3.3.2 Color Feature Extraction:

Histogram is a demonstration of sharing of color in an image. With the help of color histogram method, the images that have been matched with text and shape feature can be compared. Histogram method enables for summarization of data distribution by identifying color matching of image. The repetition of values in the range of those pixels is calculated by considering some range of pixel values. The brightness distribution for these colors can be finest which is described by the color histogram which will be more helpful to assess the clipping of the individual colors. The histogram for Red, Green, Blue color values of pill images can be individually calculated separately by taking grayscale value in X-axis and number of recurrence of that value at image in Y-axis. Together of those values get varied for every image [2]. Histogram is considered and derived for image that is well matched with database image. The color histogram is a very common approach to compare the color component of a query image to that of database images. Hence, this methodology relies on such pixel values. Color histograms are measured for each image so that to identify relative proportions of pixels within certain values. The specific color values helps for searching images from a database and it is the most basic form of color retrieval. All visible colors along with the basic combination of some set of basic color namely Red, Green and Blue (RGB) are represented by the computers [2]. Based on the similarity of three unlike histograms in which one for RGB pixel, the image retrieval measures likeness for being utilised during their experimentation. The intersection of such color histogram was already proposed for color image retrieval [8]. The intersection of histograms h and g is given by:

$$d(h,g) = \frac{\sum_A \sum_B \sum_C \min(h(a,b,c), g(a,b,c))}{\min(|h|,|g|)} \quad (8)$$

where, $|h|$ and $|g|$ represents to the magnitude of each histogram which is equal to the number of samples. The background colors are eliminated that are not present in the user's query image do not contribute to the intersection distance. The normalized sum is resulting by the histogram using few samples.

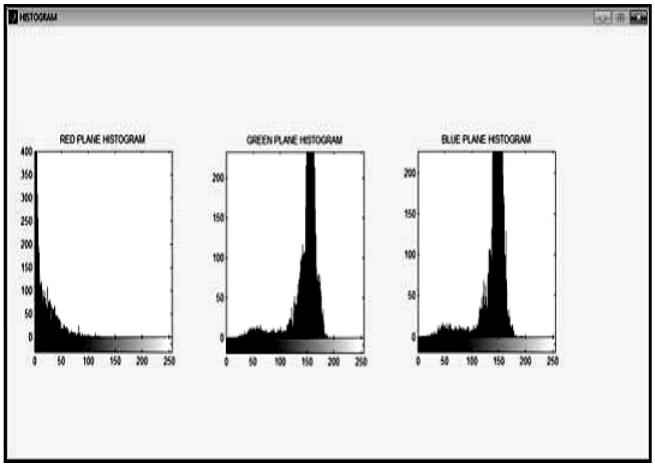


Fig.5. Color Histogram

3.4 MATCHING

In order to arrive at visually similar results, several features are involved in the matching process. It is computed by evaluating the distance of both database and query image. The distance measure motivates the use of cross-correlation for matching as given by,

$$d_{f,t}^2(u,v) = \sum_{x,y} [f(x,y) - t(x-u, y-v)]^2 \quad (9)$$

where, *f* is the image and the sum is over x,y under the window containing the feature t positioned at u,v,d^2 represents squared Euclidean distance.

The primary colors such as red, green and blue are constituted in RGB color model. The color model used in this system is CRT monitors and color raster graphics and hence defined in this system. As the colors are added further to produce the specific color, the models are considered as the "additive primaries". The RGB model employs the Cartesian coordinate system (0, 0, 0) to convert black to (1, 1, 1) white as represented by the grey-scale.

4. EXPERIMENTAL RESULTS

With the help of the pharmaceutical database, the experiment is carried out and the results are shown as follows. Using MATLAB, the outcome of the pill whether legal or illegal is identified.

The result is shown matching and referred as legal pill if the known query pill is matched with the entire three features like text, color and shape. If the pill is not matched with any one of these three features, the result will be shown as 'Pill not matched'. This mismatch occurring due to any one of the features of text, color and shape is declared as illegal pill. It is illustrated as shown in figure.

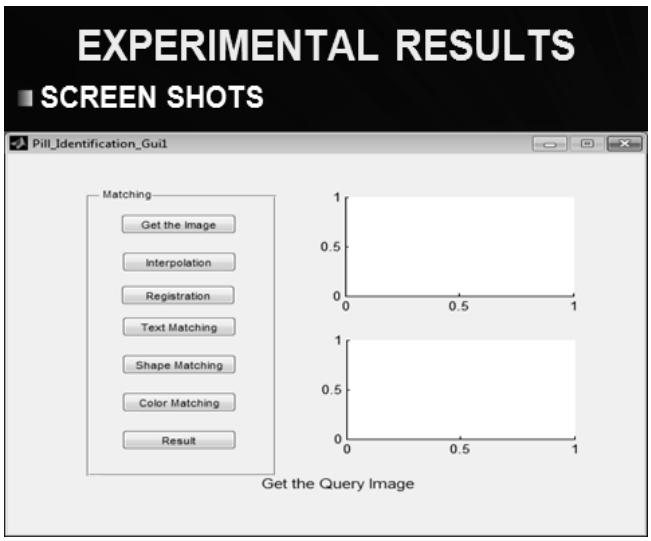


Fig.6. Result screen shot

The evaluation parameters are Sensitivity, Specificity and Accuracy has been calculated to assess the result of images. These parameters are determined for different criterion parameters such as True positive (TP), True negative (TN), False positive (FP), False negative (FN) as follows. Accuracy is also calculated in the proposed system based on the formula given below as.

Sensitivity = (TP/TP+FN)

Specificity = (TN/TN+FP)

Accuracy Percentage = ((Matched Pill)/ (Matched Pill +Miss Matched Pill))*100

Fig.7. Matched pill screen shot

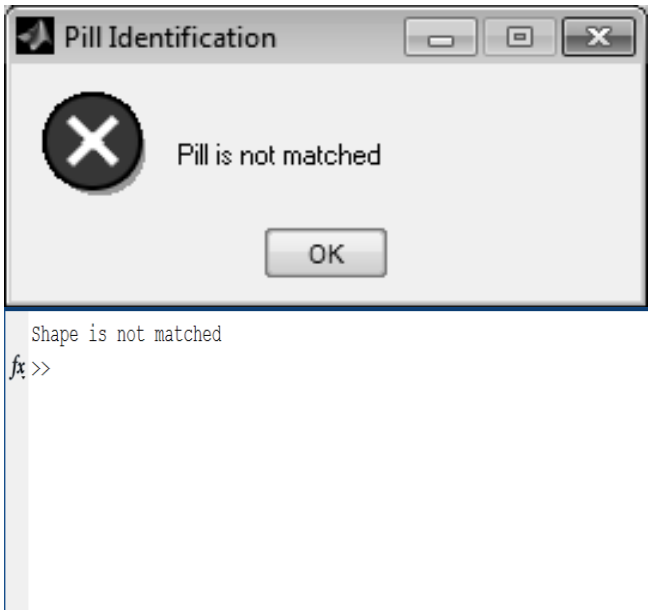


Fig.8. Screen shot of not matched pill

Table.1. Example Accuracy evaluation

Factor	Positive	Negative
True	145	35
False	16	11

Accuracy = ((TP+TN)/(TP+TN+FP+FN))^100

Accuracy Percentage = 86.76%

Table.2. Overall Accuracy calculation

System	True Positive	True Negative	False Positive	False Negative	Accuracy
Proposed	675	389	90	69	86.99918
IQDI	655	368	110	90	83.64677
IDPI	579	373	186	85	77.84137
QBIC	584	334	181	124	75.06133

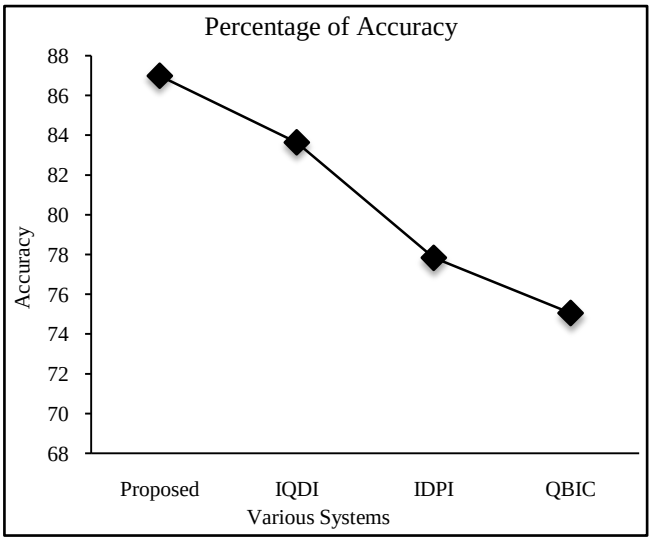


Fig.9. Overall Accuracy chart

Fig 10.Time Calculation

The average accuracy of 86.9 % was obtained based on the proposed system after matching with text, color, shape feature. The total elapsed time of 4.48 Sec was taken for the entire process and elapsed time for each individual feature is also shown in the Fig.9. Thus, the results from various experiment showed that algorithm used in the proposed system is efficient when compared to other existing algorithm.

Table.3. Tabulation for overall comparison of the parameters

Parameters	QBIC	IDPI	IQDI	Proposed Approach
Accuracy	75 %	78%	84%	87%
Sensitivity	84.1 %	86.02%	87.5 %	88.5%
Specificity	87.3 %	86.8%	84 %	82%

5. CONCLUSION

The keywords available in the existing conventional system are subjective as not all the information about the pill is captured for accurate retrieval and moreover the Artifacts, Blurring and Jagged image zooming process were resulted. Most of the existing identifier tools are keyword based where a keyword in text format is used to search a database having pill images and details. The demerit of such approaches are that the keywords are normally subjective and do not capture all the information about the pill for accurate retrieval. In certain cases, correct identification of pills become challenging due to very few or misappropriates keywords.In this research work, in order to solve the above said problems, an efficient drug identification system using image mining techniques has been proposed. Image mining is the idea used to detect unusual patterns and extract implicit and useful data from images stored in the large data bases. Therefore, it is observed that image mining deals with bringing out correlations among different images from large image databases. Simulation results shows that the proposed techniques provide accurate retrieval results both in terms of time and accuracy when compared to conventional approaches. This proposed system employed combination of techniques and hence reduced the limitations of the existing system by providing accurate retrieval results in terms of both time and accuracy when compared to existing system.

REFERENCES

[1] AGS Expert Panel, "Guiding Principles for the Care of Older Adults with Multimorbidity: An Approach for Clinicians, American Geriatrics Society Expert Panel on the Care of Older Adults with Multimorbidity", *Journal of the American Geriatrics Society*, Vol. 60, No. 10, pp.1-25, 2012.

[2] Rishav Chakravarti and Xiannong Meng, "A Study of Color Histogram Based Image Retrieval, *Sixth International Conference on Information Technology: New Generations*, pp. 1323-1328, 2009.

[3] E. Annasaro and A. Hema, "Enhanced Image mining techniques for drug pill image", *International Journal of Computer Trends and Technology*, Vol. 4, No. 2, pp. 94-102 , 2013.

[4] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions", *Bulletin of the Calcutta Mathematical Society*, Vol. 35, pp. 99-109, 1943.

[5] S. Balan and T. Devi, "Design and Development of an Algorithm for Image Clustering In Textile Image Retrieval Using Color Descriptors", *International Journal of Computer Science, Engineering and Applications*, Vol. 2, No. 3, pp. 199-211, 2012.

[6] Z. Geradts, and J. Bijhold, "Content based information retrieval in forensic image databases", *Journal of Forensic Sciences*, Vol. 47, No. 2, pp. 285-292, 2002.

[7] W. Hsu, S. Antani, L. Long, L. Neve and G. Thomas, "SPIRS: A Web-based image retrieval system for large biomedical databases", *International Journal of Medical Informatics*, Vol. 78, No. 1, pp. S13-S24, 2009.

[8] Y. B. Lee, U. Park, A. K. Jain and S. W. Lee, "Pill-ID: Matching and retrieval of drug pill images", *Pattern Recognition Letters*, Vol. 33, No. 7, pp. 904-910, 2012.

[9] M. Pietik¨Ainen, and O. Okun, "Text Extraction from Grey Scale Page Images by Simple Edge Detectors",

Proceedings of the12th Scandinavian Conference on Image Analysis, pp. 628-635, 2001.

[10] C. Schmid and R. Mohr, "Local gray value invariants for image retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 5, pp. 530-534, 1997.

[11] Shamik Sural, Gang Qian and Sakti Pramanik, "Segmentation and Histogram Generation Using The HSV color Space For Image Retrieval", *Proceedings of the IEEE International Conference on Image Processing*, Vol. 2, pp. II-589-II-592, 2002.

[12] R. Venkata Ramana Chary, D. Rajya Lakshmi and K. V. N. Sunitha, "Feature Extraction Methods For Color Image Similarity", *An International Journal on Advanced Computing*, Vol. 3, No. 2, pp. 147-157, 2012.

[13] Q. Yuan and C. L. Tan, "Text Extraction from Gray Scale Document Images Using Edge Information", *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pp. 302-306, 2001.

[14] http://en.wikipedia.org/wiki/Health_care_industry.

[15] http://www.pharmer.org/images.

[16] http://www.drugs.com