Kurdistan Journal of Applied Research (KJAR) | Print-ISSN: 2411-7684 – Electronic-ISSN: 2411-7706 | kjar.spu.edu.iq
Volume 2 | Issue 1 | June 2017 | DOI: 10.24017/science.2017.1.12

An Online Content Based Email Attachments Retrieval System

Noor Ghazi M. Jameel
Technical College of Informatics,
Sulaimani Polytechnic University,
Sulaimani, Iraq
Noor.ghazi@spu.edu.iq

Esraa Zeki Mohammed
Kirkuk Dept,
State company for Internet Services,
Kirkuk, Iraq
Isramohammed2@gmail.com

Loay Edwar George
Computer Science Dept,
University of Baghdad,
Baghdad, Iraq
loayedwar57@scbaghdad.edu.iq

Abstract: ***E-mail is one of the most popular programs used by most people today. As a result of the continuous daily use, thousands of messages are accumulated in the electronic box of most individuals, which make it difficult for them after a period of time to retrieve the attachments of these messages. Most Email providers constantly improved their search technology, but till now there is something could not be done; i.e., searching inside attachments. Some email providers like Gmail has added searching words inside attachments for some file types (.pdf files, .doc documents, .ppt presentations) but for image files this feature not supported till now. However, E-mail providers and even modern researches have not focused on retrieving the image attachments in the E-mail box. The paper was aimed to introduce a novel idea of using Content based Image Retrieval (CBIR) in E-mail application to retrieve images from email attachments based on entire contents. The work main phases are: feature extraction based on color features and connect to Email server to read Emails, the second phase is retrieving similar image attachments. The tests carried on mail inbox contain 100 messages with 500 image attachments and gave good precision and recall rates When the threshold value is less than or equal to 0.4.***

Keywords: CBIR, Color Features, Email Attachments, Email Retrieval System, Image Retrieval, Similarity Measure.

1. INTRODUCTION

The Methodology for searching images efficiently is an important research topic and retrieving images that match user's needs is not a simple task [1].

These days, images are used in numerous applications; hence, finding successful techniques for retrieving images have gotten extensive interest. To overcome the problems of the traditional approaches for retrieving images based on keywords, CBIR was introduced [2]. The most recognized feature for image retrieval is color. It considered as primitive feature for classy image retrieval systems. One of the methodologies used for color feature extraction is Color Histogram (CH). CH shows the distribution of color contents in an image. It is very fast and efficient technique. Many commercial and academic systems used CH for image retrieval such as QBIC, NETRA, RETIN, KIWI, and Image Minor [3]. Email still fill in as imperative application to store data and information for their day by day activities [4]. Some of this information is attachments attached to email messages. Attachments include images, audio, video, PDF, Word documents, and so on. In this paper an online images retrieval system is introduced to retrieve images from email attachments based on the content of the image.

2. LITERATURE REVIEW

Recently, there was noticeable increase for utilizing the developed CBIR methods in different applications, for example:

Loay and Mohammed [5], improved the retrieval performance based on texture features. They use 600 samples from variety human tissues and the results reflected very high retrieving rates.

Alsmadi and Alhami [6], evaluated several approaches to cluster emails based on their contents. For classification purpose algorithms were developed for large collection of text.

Yuvaraj and Hariharan [7], presented similar objects matching depending on three features using computer vision. The experiments were conducted using Matlab software; the results indicated that region based and color histogram based methods are effective methods.

Dubey et al. [8] introduced two multi-channel decoded local binary patterns; the experiments applied on 10 DB with variety natural scene and textures.

PyykkÖ and Glowacka [9] used deep neural network for interactive content based image retrieval by using few training samples to learn automatically from users' interaction and feedback to reduce the training time. Image features were extracted using Convolutional Neural Networks (CNN).

Parthiban and Srinivasa [10] used Adaboost algorithm to classify images based on bag of features to minimized the storage cost and for efficient retrieval.

3. CONCEPTS AND METHODS

3.1 Email System

The electronic mail is one of the most common internet services, it remains one of its important applications over the years. Email has enormous features, including sending messages with hyperlinks, attachments, HTML text, and embedded photos [11].

3.1.1 Email System Architecture

The email system architecture is illustrated in Figure (1). It contains two sub systems: (i) the user agents are used to read, send, compose, replies to messages, display incoming messages, and arrange messages by filing, searching, and deleting them. Examples to most common user agents are Google Gmail, Microsoft Outlook, Mozilla and Apple Mail. (ii) The message transfer agents, are used to send messages from the source to the destination with the help of Simple Mail Transfer Protocol (SMTP). They are also known as mail servers. [12][13].

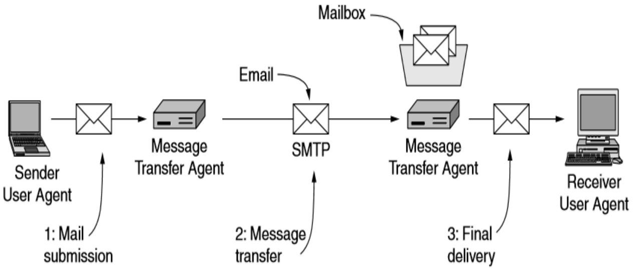
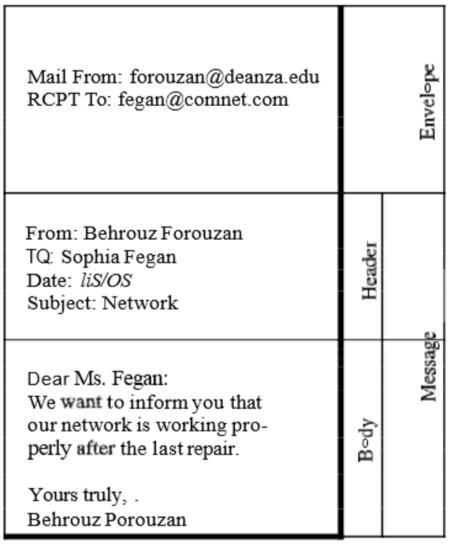


Figure 1 Architecture of email system [13]

3.1.2 Email Message Format

The email has an envelope and a message. The sender and the receiver addresses are contained in the envelope part of the email. The message part contains the header and the body. Messages must be formatted in a standard way to be handled by message transfer agents. RFC 822 is a standard format which defines messages to have a header and a body and they are represented in ASCII text. primarily, the body was supposed to be simple text. RFC 822 was updated quite a few times to allow email messages to support and transfer many different types of data: audio, video, images, PDF documents, and so on [13]. The header specifies the sender, the receiver, the subject of the message, and some other information (e.g. content type, encoding type, etc.). The body contains the actual information to be read by the receiver. The general layout of Email file is illustrated in Figure (2) [14].



Mail From: forouzan@deanza.edu
RCPT To: fegan@comnet.com

From: Behrouz Forouzan
TQ: Sophia Fegan
Date: *liS/OS*
Subject: Network

Dear Ms. Fegan:
We want to inform you that our network is working properly after the last repair.

Yours truly, .
Behrouz Porouzan

Figure 2 Electronic Email [14]

3.1.3 Search Mechanism in Email Agents

Email clients have gotten a lot smarter agents over the last 10 years, especially their search features. Many web based email services and email clients offer search mechanisms for the full text of the message, many companies offer desktop applications that can support indexing, searching the file systems, emails and the browser caches and there are also many research prototypes which perform the search operation [4]. User agents recently offer wide capabilities to search the mailbox. Search capabilities let users find messages quickly, for example message that someone sent in the last month about specific topic [13]. Gmail, yahoo, and many other email clients provide search capabilities; like search messages (From) or (To) fields for specific email addresses or people, search for keyword or word in the header or the body of the message, messages sent or received before or after specific date or in specific period of time, messages with file size, search for messages that have files attached to them, messages that are starred, unread, read or chat message, and search for file names of attachments or files with extensions .jpg, .pdf, .doc, .ppt, .xls and return emails with the specific file with that extension. But till now there is no search methodology to retrieve the image files from emails depending on the content of the image.

3.2 CBIR

In 1992, Kato [15] introduced the concept of CBIR to describe images retrieving from a database automatically by using the color and shape features [16]. The main tasks for CBIR systems are the similarity comparison that depend on finding the difference between query image feature with the corresponding features of other image stored in a database [17].

3.3 Image Histogram

In this work a conventional color histogram (CCH) used to indicate occurrence of every color in an image for representing the statistical behavior of each color in image.

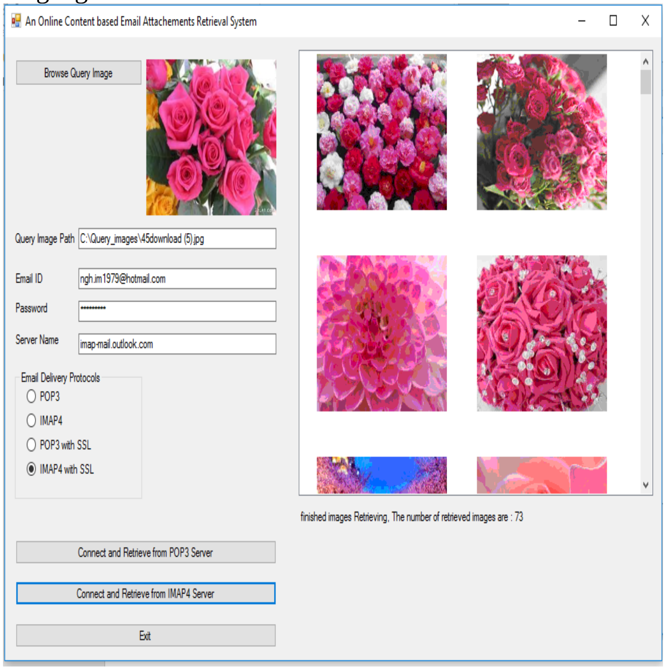
$$\Pr(i) = \frac{h_i}{m \times n} \qquad (1)$$

Where, h_i represents the number of pixels in color C_i [18].

4. THE PROPOSED SYSTEM

The proposed email attachments retrieval system is client based which is shown in Figure (3), it uses query-by-example (QBE) paradigm. An image sample based on what a user needs to search or find in email attachments loaded to the system and the similar images to a given sample are retrieved from email attachments.

First, a user starts by uploading the image sample from the main system interface and enters his Email ID, password, server name and email delivery protocol then connects to the mail server. For test purposes the system connected to a real mail server (Hotmail server). Then, the mail server will check the entered information, if it is correct, then the system will read each email from the user's mailbox. The mailbox contains email messages with and without attachments. The system will check every email if it contains attachment or no. If the email contains image attachment(s), then the color image histogram features are computed for them to use later for comparison with query image feature vector. Then set of attachment images that have high similarity to the query image are retrieved, displayed and saved in the list containing the file name (email number with the attachment file name) to avoid duplicates. The system was developed using Visual Basic.Net programming language.


An Online Content based Email Attachements Retrieval System

Query Image Path: C:\Query_images\45download (5).jpg

Email ID: ngh.im1979@hotmail.com

Password: **********

Server Name: imap-mail.outlook.com

Email Delivery Protocols
- ○ POP3
- ○ IMAP4
- ○ POP3 with SSL
- ● IMAP4 with SSL

finished images Retrieving, The number of retrieved images are : 73

Browse Query Image
Connect and Retrieve from POP3 Server
Connect and Retrieve from IMAP4 Server
Exit

Figure 3 The Interface of the Email Attachments Retrieval System

The block diagram of the email attachments system is shown in Figure (4) and explains the steps of the proposed system in general. The implementation of automated identification of attachment images illustrated in the flowchart in figure (5) and implies the following steps:

4.1 Loading Image, Read, Parse, and Check Email Attachments

This step loads the data of input image. Also, through the application, the user will enter his email ID, password, server name and email delivery protocol, then connect to the mail server. Port 110 is the default POP3 server port to receive Emails. Port 995 is the common POP3 Secure Socket Layer (SSL) port used to receive email over implicit SSL connection. Port 143 is the default IMAP4 server port, Port 993 is the common port for IMAP4 SSL. Now SSL is commonly used, many email servers require SSL connection such as Gmail, Outlook, Office 365 and Yahoo. In this system a connection to Hotmail server was done using IMAP4 through SSL and IMAP Hotmail server name (imap-mail.outlook.com). If the connection was successful, each email will be read from the mail server and parsed into header and message body. The body will be checked if it contains attachments or no. If it contains attachments, the files will be read and checked if they are images with the extensions (JPG, BMP, or GIF). The color image histogram will be computed for image attachments.

4.2 Compute Color Image Histogram

A color histogram is computed for every image used in the proposed system, the x-axis represents the number of colors in an image. The y-axis represents the number of pixels there are in each color [18].

4.3 Distance Measure

The similarity measure between Q_j (Query Image) and T_k (Attachment Image) having feature vectors $\{q_{ji}|i=0\ldots N\text{-}1\}$ and $\{T_{ki}|\ i=0 \text{ to } N\text{-}1\}$ is computed using Euclidean distance metric [19]:

$$d(Q_j, T_k) = \sum_{i=0}^{N-1} |Q_{ji} - T_{ki}| \quad , \qquad (2)$$

if the similarity is less or equal to the threshold, the attachment images will be retrieved and stored on the computer to be displayed later after the retrieval process completed and all emails were checked.

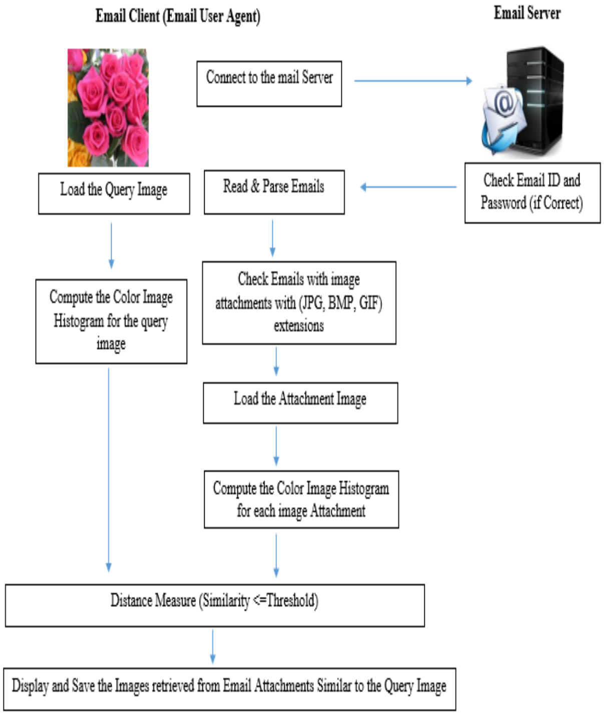

Email Client (Email User Agent)
Email Server
Connect to the mail Server
Load the Query Image
Read & Parse Emails
Check Email ID and Password (if Correct)
Compute the Color Image Histogram for the query image
Check Emails with image attachments with (JPG, BMP, GIF) extensions
Load the Attachment Image
Compute the Color Image Histogram for each image Attachment
Distance Measure (Similarity <=Threshold)
Display and Save the Images retrieved from Email Attachments Similar to the Query Image

Figure 4 The System Block Diagram

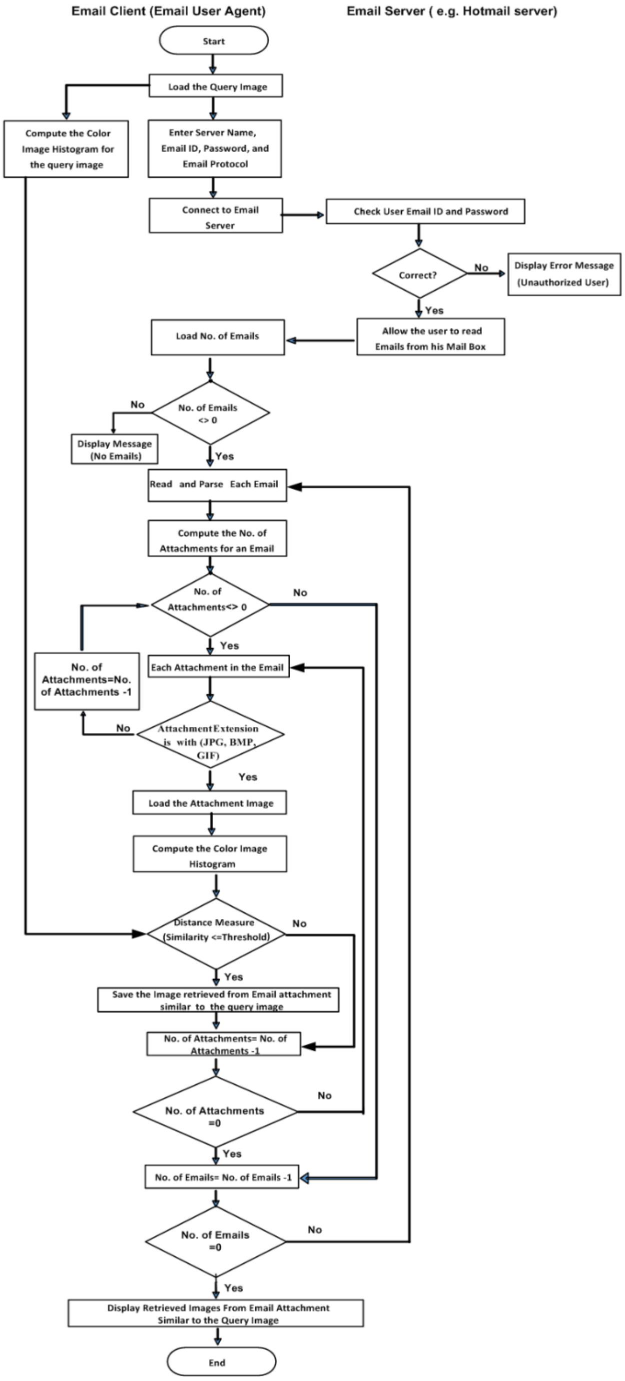

Email Client (Email User Agent)
Email Server (e.g. Hotmail server)
Start
Load the Query Image
Compute the Color Image Histogram for the query image
Enter Server Name, Email ID, Password, and Email Protocol
Connect to Email Server
Check User Email ID and Password
Correct?
No
Display Error Message (Unauthorized User)
Yes
Allow the user to read Emails from his Mail Box
Load No. of Emails
No. of Emails <> 0
No
Display Message (No Emails)
Yes
Read and Parse Each Email
Compute the No. of Attachments for an Email
No. of Attachments<> 0
No
Yes
Each Attachment in the Email
No. of Attachments=No. of Attachments -1
No
AttachmentExtension is with (JPG, BMP, GIF)
Yes
Load the Attachment Image
Compute the Color Image Histogram
Distance Measure (Similarity <=Threshold)
No
Yes
Save the Image retrieved from Email attachment similar to the query image
No. of Attachments= No. of Attachments -1
No. of Attachments =0
No
Yes
No. of Emails= No. of Emails -1
No. of Emails =0
No
Yes
Display Retrieved Images From Email Attachment Similar to the Query Image
End

Figure 5 System Flowchart

5. RESULTS AND DISCUSSION

The conducted tests results are presented in this section to show the performance of the established system whose structure is introduced above. As well as, the tests are arranged to explore the effects using different threshold values on the overall system retrieval performance.

For retrieval purpose two metrics were used; they are [20]:

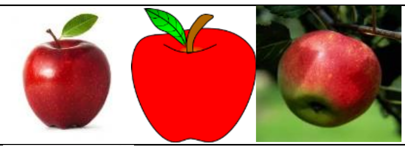
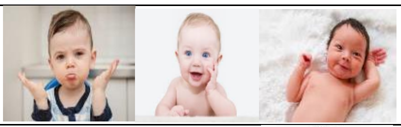
$$\text{Precision} = \frac{\text{Retrieved Related Images}}{\text{Total Retrieved Images}} \times 100\% \quad (3)$$

$$\text{Recall} = \frac{\text{Retrieved Related Images}}{\text{Total Related Images}} \times 100\%, \quad (4)$$

The data sets used in this study are sets of email attachment images with different extensions (e.g., .bmp, .jpeg, .gif) which contain different subjects (e.g., apples, cars, chairs, babies face, flowers, grass, mobiles, sea, scanned documents) of varying sizes.

About 500 images were used in this test taken from 100 email messages, as well as other set of images were used for test purpose. Table (1) presents examples of the used ten image data sets which have been used.

Table1: Examples of Images Data Set

Classes of image	Example of images	No. of images
Apple		50
Chair		35
Baby Face		15
Formal Paper		50
Flowers		100
Grass		25
Mobile		50
Red Cars		85
Sea		50
White Cars		40

One of the main concerns in the conducted tests is to find the suitable value of threshold parameter; which leads to more accurate retrieval. If the value of threshold is too small, then the number of retrieved images will greatly have decreased and only the very similar images will be retrieved. But, if the value is too large, then, images from another set may retrieved. There is no analytical method for finding the optimal threshold value; it is usually assessed using trial mechanism (i.e., trying different values tuning the system performance) as shown in Table (2).

Table2: The Effect of Distance Measure Threshold Value

Threshold Value	Apple		Chair		Baby Face	
	Preci-sion	Re-call	Preci-sion	Recall	Preci-sion	Recall
<= 0.3	73%	26%	77%	21%	84%	25%
<= 0.4	53%	31%	59%	28%	76%	29%
<= 0.5	46%	45%	46%	57%	57%	40%
<= 0.6	32%	56%	36%	58%	48%	53%
<= 0.7	28%	68%	29%	65%	32%	69%

Threshold Value	Formal Paper		Flower		Grass	
	Preci-sion	Re-call	Preci-sion	Recall	Preci-sion	Recall
<= 0.3	76%	19%	76%	30%	84%	34%
<= 0.4	66%	29%	65%	46%	76%	47%
<= 0.5	50%	32%	53%	56%	63%	55%
<= 0.6	42%	46%	43%	68%	51%	69%
<= 0.7	31%	59%	33%	79%	42%	75%

Threshold Value	Mobile		Red Cars		Sea	
	Preci-sion	Re-call	Preci-sion	Recall	Preci-sion	Recall
<= 0.3	95%	21%	81%	30%	85%	21%
<= 0.4	83%	34%	60%	45%	74%	32%
<= 0.5	75%	48%	49%	56%	55%	36%
<= 0.6	63%	59%	33%	64%	49%	53%
<= 0.7	51%	71%	30%	73%	38%	70%

Threshold Value	White Cars	
	Prec-ision	Recall
<= 0.3	88%	15%
<= 0.4	78%	22%
<= 0.5	59%	34%
<= 0.6	48%	51%
<= 0.7	35%	71%

Figure (6) illustrates the effect of different values for threshold parameter on the precision for each image category. Figure (7) shows the effect of different threshold parameter values on the recall for each image category.

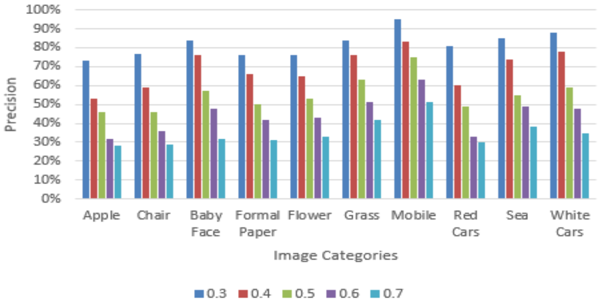


Figure 6 The Effect of Threshold on Precision

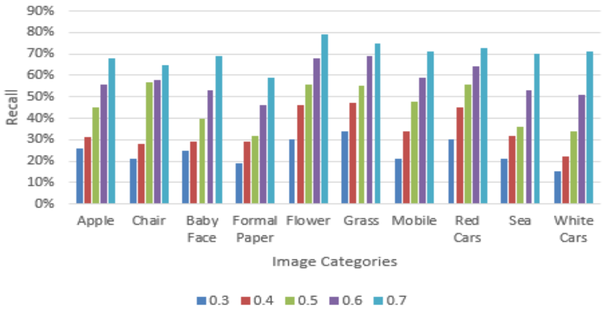


Figure 7 The Effect of Threshold on Recall

6. CONCLUSION

The proposed retrieval system facilitates access to email attachments images in mailbox based on interaction user interface that allow user to quickly obtain an overview of similar images in email account. The color histogram can be used to describe the color content of images. Testing the different threshold values helps for best retrieval results. The system gave better rates, when the threshold value is less than or equal (0.4).

7. REFERENCE

1. X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, “Enhancing sketch-based image retrieval by re-ranking and relevance feedback”, IEEE Trans. Image Processing, Vol. 25, pp. 195-208, 2015.
2. M. Azodinia, and A. Hajdu, “A Novel combinational relevance feedback based method for content-based image retrieval”, ActaPolytechnicaHungarica, Vol. 13, no. 5, pp. 121-134, 2016.
3. A. Saini and R. Bharti “A review on content based image retrieval by different techniques”, International Journal of Neural Systems Engineering, Vol. 1, no. 1, pp. 1-6, 2017
4. S. B. Pitla, “Organizational Search in Email Systems”, M.S. thesis, Dept. Mathematics and Computer Science, Western Kentucky Univ., 2012.
5. L. E. George, and E. Z. Mohammed, "Tissues image retrieval system based on Co-occurrence, run length and roughness features", IEEE Conference Publications, International Conference on Computer Medical Applications (ICCMA), DOI: 10.1109/ICCMA.2013.6506186, pp. 1-6, 2013.
6. I. Alsmadi, and I. Alhami, “Clustering and classification of email contents”, Journal of King Saud University – Computer and Information Sciences, Production and hosting by Elsevier B.V. on behalf of King Saud University, Vol. 27, pp. 46–57, 2015.
7. D. Yuvaraj, and S. Hariharan, “Content-based image retrieval based on integrating region segmentation and colour histogram”, International Arab Journal of Information Technology, Vol. 13, pp. 203-207, 2016.
8. S. R. Dubey, S. K. Singh, and R. K. Singh, “Multichannel decoded local binary patterns for

content based image retrieval", IEEE Trans. Image Processing, Vol. 25, pp. 4018-4032, 2016.

9. J. PyykkÖ and D. Glowacka, “Interactive content-based image retrieval with deep neural networks”, Symbiotic 2016, LNCS 9961, pp. 77–88, 2017
10. Parthiban S. and Srinivasa Raghavan S., “Content based image classification and retrieval using visual bag of features and adaboost algorithm”, ARPN Journal of Engineering and Applied Sciences, Vol. 12, No. 2, pp. 588-590, 2017.
11. J. F. Kurose, and K. W. Ross, “Application layer in Computer Networking a Top-Down Approach”, 6th ed., USA: Pearson Education, Inc., pp. 118-130, 2013.
12. A. S. Tanenbaum, and D. J. Wetherall, “The application layer in Computer Networks”, 5th ed., USA: Pearson Education, Inc., pp. 623-646, 2011.
13. L. L. Peterson and B. S. Davie, “Application in Computer Networks a systems approach”, 5th ed., USA: Elsevier, Inc., pp. 700-708, 2012.
14. B. A. Forouzan, “Remote logging, electronic mail, and file transfer” in “Data Communications and Networking”, 4th ed., USA: McGraw-Hill, pp. 824-840, 2007.
15. T. Kato, "Database Architecture for Content-Based Image Retrieval", Proceedings of Image Storage and Retrieval Systems (SPIE), pp. 112-123, 1992.
16. J. Eakins, and M.Graham, "Content-based image retrieval", University of Northumbria at Newcastle, Report no. 39, 1999.
17. E. Aulia, "Hierarch Indexing for Region Based Image Retrieval", M.Sc. Thesis, Department of Industrial and Manufacturing Systems Engineering, Louisiana State University, 2001.
18. J. Huang, "Color-Spatial Image Indexing and Applications", Ph.D. Thesis, Cornell University, 1998.
19. C., Li Wei, C., and R.Wilson, "A general framework for content-based medical image retrieval with its application to Mammograms", Proceedings of the SPIE, Vol. 5748, pp. 134-143, 2005.
20. G.Brunner, "Structure features for content-based image retrieval and classification problems", Ph.D. Thesis, University of Freiburg, Germany, 2006.