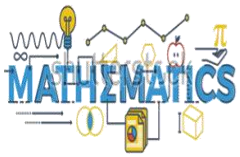
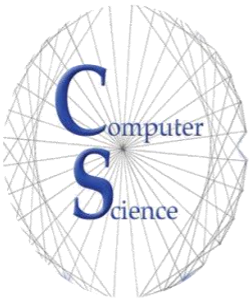
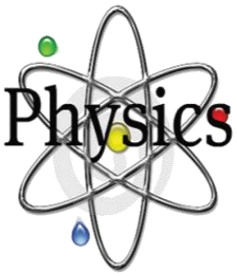
Part 3

Computer Scope







Content Table

A Prediction of Grain Yield Based on Hybrid Intelligent Algorithm

Ibrahim Ahmed Saleh[1], Wasan Abdallah Alawsi[2] , Omar Ibrahim Alsaif [3] , Khalil Alsaif[4]

1University Of Mosul, Software Engineering Dept. Iraq, i.hadedi@uomosul.edu.iq
2 University Of Al- Qadisiyah, college Science, Iraq, wasan.alawsi@qu.edu.iq
3 Northern Technical University, Mosul Technical Institute, Iraq, omar.alsaif@ntu.edu.iq
4 University of Mosul, Computer Science Dept., Iraq, khalil_alsaif@hotmail.com

Abstract: The prediction is most important goals in economic quantitative studies, it basis in design and plan future economic policies properly process over forecasting accuracy. This paper is aiming at the problem salp swarm algorithm (SSA) for predicting grain yield is prone to fall into the local optimal problem. An improved SSA is proposed with combine with back propagation neural network. Using the different advantages of SSA algorithm in global search capabilities, combining the two for further optimize the weight, improve the accuracy and robustness of the grain yield prediction model. The specific implementation is selected from 1963 to 2013.
These methods are used to define agricultural datasets that supports crop growth decision for grain product and its influencing factors were tested as a data set.
The results show that, the improved salp swarm optimization can be classified as a good predict tool for the domestic food production trend in recent years compared with the SSA. This paper briefly introduces three artificial methods BP neural networks, SSA and improved SSA optimization algorithm. The natural behavior of salp, barrel-shaped plankton that are mostly water by weight optimization and combined with mixed-group of intelligent algorithm are simulated. The simulation results of grain production prediction illustrate that the predict precision of the improved SSA is much higher than of both conventional BPNN and SSA techniques and it's very efficient and practicable.
Keywords: grain yield; prediction; Salp swarm optimization; neural network, back propagation algorithm.

Introduction

Grain is given the high position in the Iraqi food basket, which in turn is formed 35% of the per capita income for Iraq yield, it is one of the basic guarantees for human survival. Food is related to the country's economic prosperity and social harmony [1]. In order to ensure the sustainable development of the agricultural market, determine the complex relationship between food production and its influencing factors. The effectively grasp and regulate the influential factors that can be intervened, and establish a feasible, accurate and easy-to-realize food production dynamics Prediction models are imperative [2]. Due to the characteristics of randomness, non-linearity and dynamic nature of the food system, it is difficult for traditional prediction methods to effectively grasp and predict it. As one of the most successful optimization method currently applied, the outstanding advantage of these methods they have optimal solution for nonlinear mapping and are often used to solve complex nonlinear problems [3]. SSA was proposed to solve different kinds of optimization issues and provide various solutions to different problems [4].
Literature [5], summarized such improvements in feed-forward neural networks in more than 20 years. As a new type of biological evolution algorithm, particle swarm algorithm relies on the information exchange mechanism between its population particles and has good global optimization ability. This feature can be used for the shortcomings of BP neural network. However, particle swarm algorithm is easy to apply for the optimization process that specify local convergence and precocity [6].
Literature [7], summarized for improved back propagation neural network with particle swarm optimization (PSO) by introducing mutation operation and adaptive adjust of inertia weight, the problem of easy to fall into local optimum, premature, low precision and low later iteration efficiency of PSO are solved. The results of grain production prediction show that the predict accuracy the method is effective and feasible.
While literature [8], used Artificial Bee Colony with weighted based Fuzzy Clustering algorithm to predict a district for them country. The researchers are cropped prediction analysis consider as high

yield area and used many factors . They improved hybrid artificial bee colony with weighted based fuzzy clustering algorithm yields better performance than other clustering algorithm like k-means and k-medoids with high accuracy.

In this paper three methods for solve this problem are applied, the back propagation neural network (BPNN), salp swarm algorithm and combining between two algorithms. The proposed algorthm called improved salp swarm algorithm or hybrid swarm intelligent algorithm. This improved algorithm could describe as follows [9].

When the salp swarm algorithm falls into convergence, it can basically cover all extreme points during the establishment of the BP neural network prediction model, making its fitting ability stronger.

Salp Swarm Algorithm (SSA)

The SSA is a new nature-inspired optimizer using a method to simulate the swarm behavior of salps in nature. SSA algorithm detects, satisfies verify and strengthening tendencies that made it interesting for developing ELM training tasks. The SSA can be looked as a flexible, capable, simple, and understudied easy [8][9].

Local convergence optimal is very critical, so the salps algorithm are updated their position vectors gradually considering dynamic crowd of agents for other salps. The salp swarm optimization has iterative natural, that lead to save managing other members of swarm for better areas. The salps vector contain leader and followers which should be update their location vectors. The leader of swarm will determine the direction of food source (F) while the movement of all followers will be moved towards leader directly or indirectly [8] [10]. Figure(1) explains the salp chain movement.

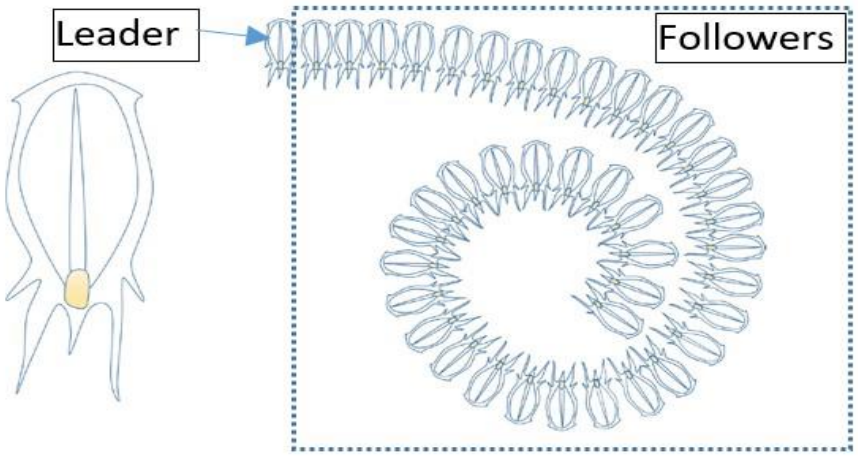


Figure 1. salp's chain movement and the model of leader and follower.

The salps population (X) contain N agents with d-dimensions, hence the matrix (N*d) can illustrated in equation (1):

$$X_i = \begin{bmatrix} x_{(1,1)} & x_{(1,2)} & \cdot & \cdot & \cdot & x_{(1,d)} \\ x_{(2,1)} & x_{(2,2)} & \cdot & \cdot & \cdot & x_{(2,d)} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot\cdot\cdot\cdot & & \cdot \\ x_{(1,N)} & x_{(2,N)} & \cdot & \cdot & \cdot & x_{(N,d)} \end{bmatrix} \qquad (1)$$

For this algorithm, the leader and followers are updated there location based on equation (2), [7]:

$$x_j^1 = \begin{cases} F_j + c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) & c_3 \geq 0.5 \\ F_j - c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) & c_3 \geq 0.5 \end{cases} \qquad (2)$$

Where:

x1j denotes the leaders 'position.

Fj represent to best solution of food source in the j th dimension
ubj , lbj are the upper and lower bounds in the jth dimension.
(c1, c2 and c3) are random values.
c1 is the main parameter of algorithm , which have significant role for the performance of SSA, It is consider as the only parameter that that calibrate and manages the balance between exploration and exploitation processes. This parameter depends on the iteration number that permits high exploration proportions as in equation (3) [9][10]:

$$c_1 = 2e^{-(\frac{4t}{T_{max}})^2} \qquad (3)$$

Where:
t iteration, while (Tmax) maximum number of iterations. By increasing iteration count, this parameter decreases. c2 and c3 generated in the period [0, 1]. As a result, it can manage to put more approve on the variegation tendency on primary steps and put more confirm on intensity movement in last periods of optimization. The followers positions can be corrected using equation (4) [11]:

$$x_j^i = \frac{x_j^i + x_j^{i-1}}{2} \qquad \ldots\ldots\ldots\ldots (4)$$

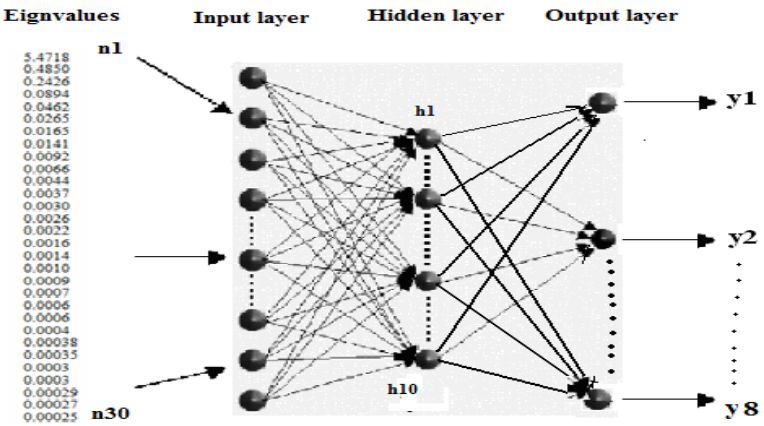
Where i ≥ 2 and xij is the location of the i th follower salp at j th dimension. The pseudo-code of SSA is expressed in Algorithm 1 as shown in Figure (2).

Step 1: Intilizaize of salps randomly $x_i=(i=1,2,\ldots,n)$
considering ***ub*** and ***lb***
Step 2: while (end condition is not met)
obtain the fitness of all salp
set *F* as leader salp
update c_1 by Eq 3
For each salps (x_i) in the population do
{
If i==1 then
Update the position of leader by Eq.2
else
Update the position of leader by Eq.4
update the population of salps based on upper and lower bounds of variables.
Update F
}
Step 3: Return F
END

Figure 2. Pseudo code of Salp Swarm Algorithm

Proposed Approach
In this paper, a crop dataset was collected from which included the Ministry of Agriculture – Department Planning and follow-up, Agricultural Statistics and Manpower Division, and - Supply Card Department. The dataset is preprocessing by redundancy by particle data filtering, and three algorithms were applied using Matlab software to predict which area will produce high yields for specific crop in particular season.
3.1 Back Propagation Neural Network Algorithm
(BPNN) is a multi-layer feed forward neural network, proposed since 1986, So far its application has spread to various fields for arbitrary nonlinear programming, BPNN can effectively fit and predict. It's mainly divided into two behaviors: forward transmission of the input signal and error reverse transmission of signals [10].

During this period, the network model tends to be optimized. The main steps in establishing a BP neural network are weights and thresholds. Optimization is occurred through training on a large number of data sets establishes and generates a network model, as shown in figure (3).



Here, X1, X2 ,… , Y n are the predictive values, w ij and wjk weights between input- hidden layer and hidden layer-output layer respectively the training process as shown in figure (4)[11].

Step 1: set the initialize network parameters (numbers of input layer nodes, hidden layer nodes and output layer nodes) and initial weights.

Step2: Calculate the hidden layer output based

$$H_j = f\left(\sum_{i=1}^{n}(w_{ij}\,x_i - a_i\right) \quad j =, 2 \ldots l$$

Where: a_j is the threshold value equal and
***f*: activation function of hidden layer**

Step 3: Calculate the output value of output layer

$$O_k = \sum_{j=1}^{l} H_j\,w_{jk} - b_k \quad k = 1, 2, \ldots \ldots m$$

Where, b_k is the threshold value of output layer node.

Step 4: Calculate the prediction error according to the network predicted output and the desired output.

$$e_k = Y_k - O_k$$

Step 5: Update the connection weights by the prediction error e_k.

$$w_{ij} + 1 = w_{ij} + \eta\, H_j(1 - H_j)x(i)\sum_{k=1}^{m} w_{jk}\,e_k \quad i = 1, 2 \ldots\ldots, n; j = 1, 2, \ldots\ldots l$$

Where, η is the learning rate.

Step 6: Update the threshold value for following questions

$$a_j + 1 = a_j + \eta H_j(1 - H_j)\sum_{k=1}^{m} w_{jk}\,e_k$$

$$b_k + 1 = b_k + e_k \quad k = 1, 2, \ldots \ldots m$$

Step 7: Determine whether the iterative ends, and if not, return to Step 2.

Figure 4. Back rogation neural network algorithm

Intermediate unit is usually set to a single hidden layer, but when the input layer have many elements, the multiple set hidden layers will be a good choice for fitting nonlinear functions better performance, more accurate predictions, and longer training time. Therefore, for more complex mapping relationships

can use multiple implicit ways to build a network. In this experiment, the network structure is more accurate in predicting food production, but it is also corresponding increased runtime.

3.2 SSA for Grain prediction

There are some essential steps should be followed to predict by SSA behavior:
Initial parameter for SSA number of swarm.
(N)=200, dim=5, lb=0, ub=1, c2=c3=0.5 , Tmax=500;
Load the data set generation values of the algorithm.
The fitness computation process which in Euclidian distance is carried out for each site visited by a (SSA) using clustering algorithm implemented for the crop dataset to predict the district as shown below.

$$D_{ij}^2 = \sum_{v=1}^{n} (X_{vi} - X_{vj})^2 \qquad (5)$$

Amid distance of the salps based on upper and lower bound of variables.
Locations update step for the exploration and centroid.

3.3. Technique of Hybrid Optimization Algorithm

The main problem of swarm algorithm lies in local or global convergence. In this paper (SSA) used to improve the inertia weight mainly optimize from two aspects. Using particles always maintain a good search ability which has fallen into convergence for points. A new method of assigning positions to make them free from the constraints of extreme points.
In the specific optimization process of SSA algorithm the weight (w) is improved with obvious effect and need to adjust quickly and effectively [12]. A larger weight can ensure that the SSA algorithm has strong search capabilities, it can make the entire algorithm quick converge, but it is easy to find optimum values [11] [12].
Combining the current research on the inertia weight of the SSA proposed a strategy to change the inertia weight following the position vector that illustrated in equation (6) [9].

$$w = w_{max} - \frac{|x_0 - g_i|}{x_{max} - x_{min}} \times (w_{max} - w_{min}) \qquad (6)$$

Where: wmax and wmin are the maximum and maximum values of the inertia weight respectively. Small values of xmax and xmin are the boundary values of particles in space, while x0 and gi the initial position of the ith particle and the optimal position of the current group. During the movement of swarm, the salps are gradually approaching the target value, then (x0-Gi) gradually becomes larger. According to equation (5), (w) shows a trend of linear decrease. Simulation tests prove that this method relatively speeds up the algorithm's convergence speed and prediction accuracy, which lead to improving the efficiency of the SSA algorithm. Predictive model of BP neural network optimized by SSA can be described as the following steps [14] [15]:

1) Initialize the parameters of the BP neural network.
2) Initialize SSA optimization algorithm to determine the population number and initial value.
3) Load the data, then (BP) neural network makes iteration over the data set generation, then returned error function is used as the adaptive value of the optimization algorithm.
4) The optimization algorithm updates the particle position according to the fitness value illustrated in equation (5) and will update after the value is returned to the BP neural network as the weight of the next iteration and threshold.
5) Substituting the weights and thresholds obtained in step (4) into step (3), Repeat iteration to the end condition of the optimization algorithm.
6) Return the final particle position to the BP neural network as the initial value of prediction model.
7) BP iterates the optimal network structure.
8) Substitute test data and output prediction results.

4. Experiment Results and Discussion

Training models in food production prediction require a lot of training data. In order to obtain better prediction results. The grain production data from 1963 to 2013 and eight factors that affect production used as data sets. The data comes from the website Iraqi ministry of agriculture (planning department), department of agricultural statistics and the central Agency for statistics and information technology, after comprehensive consideration for many factors affect to grain yield include [13] [14]:

-Sown area with grain.
- Total irrigated agricultural area.
- Manpower.
- Total amount of agricultural fertilizers.
- Total power of agricultural machinery owned at end of the year.
- Total areas guaranteed rain.
- Total rainfall areas and amount of rain.

-The total agricultural area is not guaranteed rain.

A total of eight indicators were used as the input values for the affected area, after causing incomplete function mapping, a total of (50 × 8) data input matrices and (141 × 1) data output matrices are formed. In addition, due to the unequal data acquisition methods, the results predicted and analyzed by the network model generated by the above data may not be consistent with the base and economic information. In the next few years, larger data set needs to be established for fitting prediction. For all data samples, consider the impact of noise, abnormal data, or inconsistent information acquisition methods. Before the neural network training, perform dimensionality normalization processing, so that all sample data is converted into [0, 1] Standard data [14]. An improved maximum and minimum method is used to process it, and the calculation formula is as follow:

$$x_i = 0.1 + 0.9 \times \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad i \in [1, n] \qquad (7)$$

The results prove that the normalized data has a stronger prediction effect, and avoids the phenomenon that a relatively consistent data dimension exists in the sample and returns to zero.

4.1 Prediction results and comparison processes

4.1.1 Prediction of BP neural network

When BP neural network is used to establish a prediction model after number of training when the number of network nodes is taken, the BP neural network has the best fitting ability to establish a prediction model. The samples with data numbers (1 to 121) are used as the training set, and the remaining 20 samples are used as the test set. The first (8) items of the sample will be input to the network, and the last one is output from the BP neural network. The prediction result and actual grain of BP network is shown in figure (4).

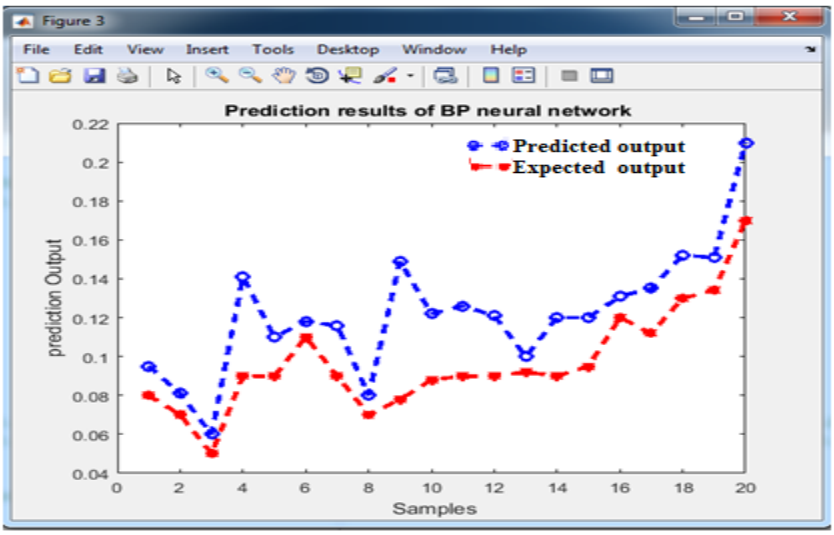


Figure 5. Prediction results of BP neural network

It can be seen from figure (5), that the predicted output results generally conform to the overall trend of the expected output, but the errors between the predicted value and the expected value are large at some points. The reasons are as follows:

1) The BP neural network is trapped in local convergence during the training process, and the best weights and thresholds are not found, so the result is a model that is not accurate enough.
2) In the test set, the mutual influence relationship between all input data is more complicated, and there are many factors influencing the grain output itself. There are large differences in the effects of natural factors and scientific and technological factors on the grain output, and the relationship between such data was not handled well when the data set was established.
3) The use of neural networks to make predictions requires a large amount of data as test samples. Under the current conditions, a large number of complex samples have not been formed, resulting in large fluctuations in some data during prediction. For example, over the years, the food disaster situation and the arable land disaster area are prone to under fitting.

4.2.2 Prediction of SSA algorithm

Salp Swarm Optimization is based on K-means clustering to analysis the district wise crop prediction, it can be seen from figure (6) the results of this algorithm.

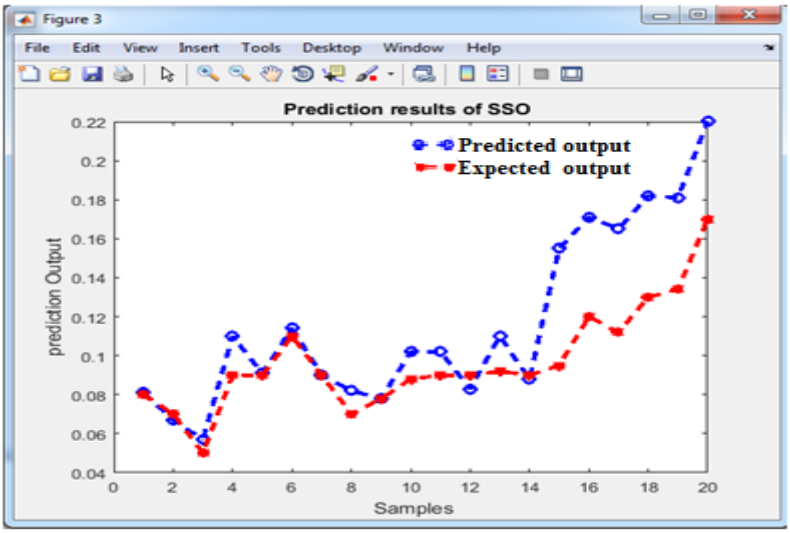


Figure 6. Prediction results of SSA algorithm

The predicted output optimized by the SSA algorithm is significantly better than the result of figure (3), but the predicted and expected values at some points are still obvious error value, and the fluctuation of the overall prediction result is large. This is because once SSA finds a better optimal value in the optimization process; other particles will quickly converge to this point.

4.2.3 Prediction of improved BP neural network

It can be seen from figure (7), that the prediction effect of improved BP neural network optimized by SSA is significantly better than the previous two prediction models, and the prediction is more accurate. The overall prediction result is consistent with the line chart trend of the expected result. The prediction effect of large fluctuation data is obviously better than the basic BP and SSA models.

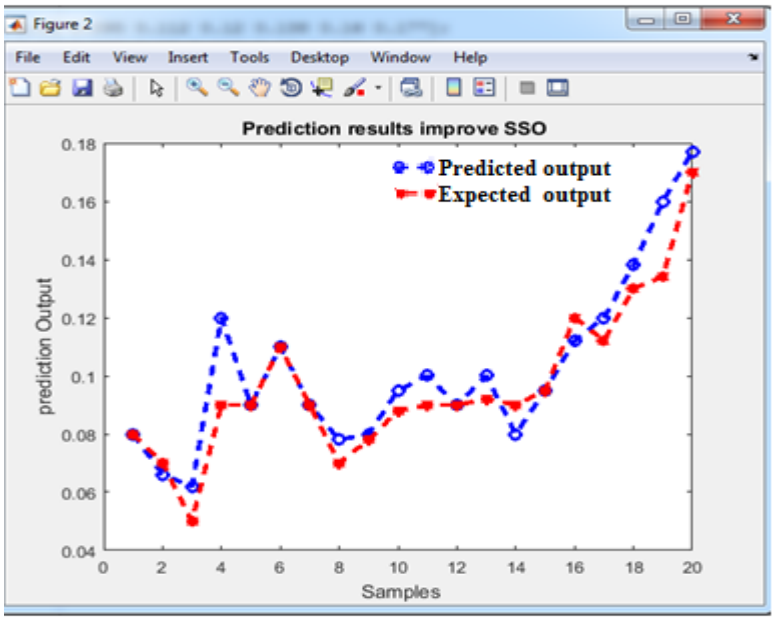


Figure 7. Prediction results of improve SSA

The prediction results of the above three methods are compared. The prediction error (the absolute value of the difference between the predicted value and the expected value) is used as the standard for the accuracy. The comparison results are shown in table (1).

Table 1. Comparison table of prediction errors of BP neural network

Network structure	maximum error	Average error (%)
Bp	18.9	8.48 ± 0.17
SSA	14.7	4.53 ± 0.15
Improve-SSA	5.88	2.42 ± 0.096

From Table 1, it can be confirmed that the improved SSA is obviously more accurate in the prediction of food production. At the same time, the optimization ability of SSA is significantly stronger than the basic Bp. The BP neural network optimized by SSA give non-linear model which has better fitting ability, less volatility of prediction results, stronger ability to resist local convergence, and achieve the expected effect.

Conclusions

At all agricultural statistical issues it is very difficult to accurately predict grain yield production, in this paper three algorithms are suggested.

1) The inertia weight update method in the SSA algorithm has been improved. The improved hybrid algorithm of SSA has strong global search performance and the ability to resist local convergence. It has played a very important role in the optimization of BP neural network weights and thresholds.
2) A feasible grain production forecasting model is established by using the existing grain production data. Through this model, the grain production can be predicted and analyzed for the next few years.
3) There are many factors that affect the grain production. The forecasting model can be used to perform some analysis one by one to determine the size of each factor's influencing factor. so the training data is improved to enhance the accuracy of the forecast.
4) The improved BP algorithm has a higher time complexity when training the model. In the future work, the search operator can be improved to enhance the search efficiency of the algorithm.

6. **References**

[1] Shaibu A. S.*, Adnan A. A. and Umar I. R. "Predicting grain yield of maize using drought tolerance traits", African Journal of Agricultural Research Vol. 10(33), pp. 3332-3337, 13 August, 2015 , DOI: 10.5897/AJAR2015.9561 . 2015

[2] Ibrahim RA, Oliva D, Ewees AA, Lu S. "Feature selection based on improved runner-root algorithm using chaotic singer map and opposition-based learning". International conference on neural information processing, Springer, Berlin, pp 156–166. 2017
[3] Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM, Salp swarm algorithm: "a bio-inspired optimizer for engineering design problems". Adv Eng Softw 114:163–191. 2017.
[4] El-Fergany AA. "Extracting optimal parameters of poem fuel cells using salp swarm optimizer". Renew Energy 119:641–648. (2018).
[5] Yok Hack, Abraham A, SON, ELV. "Meta- heuristic disaffects feed for world neural networks": over view decades for research [J]. Engineering Application Artifact Intelligence, 60: 97-116. 2017.
[6] Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., & Mirjalili, "Salp swarm algorithm: A bio-inspired optimizer for engineering design problems". Advances in Engineering Software. (2017), Volume 114, December 2017, Pages 163-191, https://doi.org/10.1016/j.advengsoft.2017.07.002.
[7] Zhang Liguo, Liu Jiangtao, and Zhi Lifu," Research on Grain Yield Prediction Method Based on Improved PSO-BP" TELKOMNIKA Indonesian Journal of Electrical Engineering Vol. 12, No. 10, October 2014, pp. 7404 ~ 7411 DOI: 10.11591 /telkomnika. v12i8.5369.
[8] Surya P. and Aroquiaraj I.Laurence, "Crop Prediction Analysis in North western Zone of Tamilnadu using Artificial Bee Colony with Weighted based Fuzzy Clustering" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6, August 2019
[9] Abusnaina Ahmed A, Ahmad Sobhi, Jarrar Radi. And Majdi Mafarja "Training Neural Networks Using Salp Swarm Algorithm for Pattern Classification", ICFNDS 2018, June 2018, https://doi.org/10.475/123_4.
[10] Ibrahim A. Saleh, Omar. I. Alsaif, SA Muhamed, EI Essa " Task Scheduling for cloud computing Based on Firefly Algorithm" Journal of Physics: Conference Series 1294 (4), 042004, 2019.
[11] Li Heli, Wang Miao, Li Bo. "Integrated particle swarm optimization for single-objective optimization Law". [J]. Journal of Chongqing University of Posts and Telecommunications. 2018.
[12] Wu Yuming, Li Jianxia. "Non-linear artificial neural network model and its application in corn production prediction". Journal of Henan normal university (Natural Science). 30: 35-38. 2002.
[13] EBERHARTR C, SHI. Y, 2000, Comparing inertia weights and constriction factors in particle swarm optimization, Proceeding of the 2000 Congress on Evolutionary Computation . Piscataway: IEEE computer Society 2000:84-88.
[14] YGOLOVN, R? NNB, CKL. "Big data normalization for massive parallel proceeding data basses". J. Computer Standards and Interfaces, , 54: 86-93. 2017.
[15] Eman A. Abdullah, Ibrahim A. Saleh, Omar I. Al Saif. "Performance Evaluation of Parallel Particle Swarm Optimization for Multicore Environment". International Conference on Advanced Science and Engineering (ICOASE) – Duhok, Kurdistan Region – Iraq. 2018

Securing Hill encrypted information with Audio Steganography: a New Substitution Method

Enas Wahab Abood[1] Wafaa A. Khudier[1] Rasha Hadi Jabber[1] Dina Adnan Abbas[2]

[1]University of Basrah-college of Science -Math. Dept
[2]Basra Education Directorate
E_mail: enaswahab223@gmail.com

Abstract: Securing data is an essential matter in communication systems, data must be protected from interceptors and eavesdroppers. In this paper, a hybrid system was produced to secure a plain text message encrypted with the Hill encryption method and embedded within an audio file in a random distribution using audio symbol sign to represent message bits. The audio file is restricted to be *.wav stereo file, Two secret keys are needed for this system encryption key and the seed of generating hiding positions. PSNR,MSE,SSIM are calculated between cover before and after embedding and the results reflected the imperceptibility requirement of the system. Also the elapsed time for securing messages was quite low for different sizes of text and the encryption process was less time than hiding.

1. Introduction

Securing data and information of all media is an essential issue that taking enormous attention and studies by researchers and especially for data transmitted over the Internet. These data pass through the net therefor it exposes the risk of third parties to interception [1]. The process of avoiding interception is difficult but, the data still need to be protected either by making it unreadable through encryption or be invisible by hiding it in Innocent other data file (steganography) [2], or even better if both methods have combined in one system [3]. Kerckhoffs defined the principle of cryptography: the cryptographic system quality should only depend on a small but significant part of the information, called a secret key. The same principle is valid for strong steganographic systems: the system knowledge that is used, must not provide any information around the existence of hidden messages [4][5][6]. Audio steganography is the science of hiding a secret message within an audio file called cover file, The secret messages can be any form of data like image, video, text or audio [7][8]. The most known and easy method for steganography is LSB method, It is a technique that allows the secret data to be inserted in the least significant bit of the audio samples of a wav file [9][10][11]. The combination of two techniques empowers the security system and increases its immunity against attackers. Only the authorized person who knows the secret keys for both techniques able to retrieve plain messages. The main goal of this paper is to produce a system that consists of two steps; the first step is cryptography to secure text messages and duplicate the security level by hiding it in a cover file of wav signal using the signal of audio file value with a random distribution. The audio file data are real values in the range [-1,1], So the values of the wav samples(positive or negative) are taken to hide message bits. A secret text message is encrypted with the Hill encrypted method then embedded within an audio file in a random distribution depending on the sign (negative or positive) of the audio values. The audio file is restricted to be *.wav stereo file, Two secret keys are needed encryption key and the seed of generating hiding positions.

2. Related Work

Audio steganography works were presented by many researchers to discuss the security of information and data file of different forms like text, images, and audio by hiding them within an audio file or the audio embedded in it. The researchers concerned with the combination of encryption and steganography. Many studies take the audio file as a matter of interest either for its importance as a file of big data to be used as a carrier file or secret information needed for protection[10]. Both frequency domain like DWT and DFT or time domain for an audio file are used for hiding, In time domain a Least significant Bit (LSB) is an effective method that could be used to insert the information in a cover file[12]. Researchers like Krishna Bhowal et.al. (2011) produced an audio Steganography for securing encrypted text messages based on and LSB technique. In their system, they used the encryption algorithm (RSA) to

encrypt the message as the first level of security then, the encrypted message bits are hidden in random and higher LSB layers that gave their system more robustness against noise addition and reduce distortion[13]. S.S. Divya and M. Ram Mohan Reddy(2012) proposed two approaches of replacement techniques for audio steganography that increases the capacity of cover audio for hiding additional data. In these methods, the message bits are hidden into a variable and multiple 4 LSBs. They used the RSA algorithm for the encryption process[14].

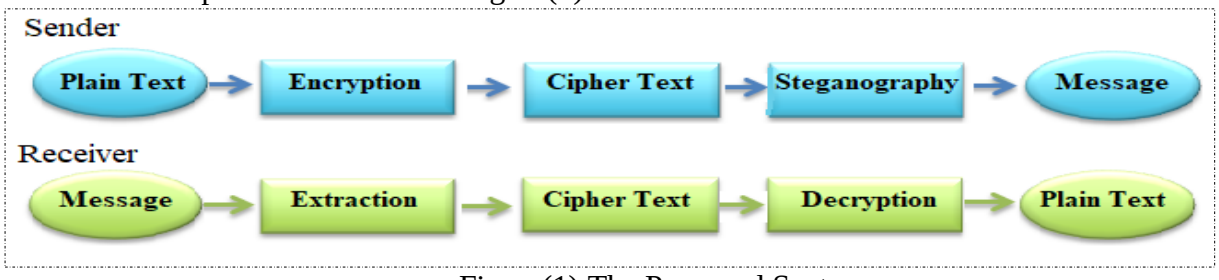
S.Hemalatha et.al. (2015) proposed an audio steganography system based on Discrete Wavelet Transform (DWT) for the cover audio file to secure data hiding. The data was an encrypted text message using a dynamic encryption algorithm. The cipher text is then hidden in wavelet coefficients of the cover audio signal[15].

Mazhar Tayel et.al. (2016) used the LSB method for embedding data in an audio file and stated that the sound quality of the cover file depends on its size and length of the hidden message [16]. Samah M. H. Alwahbani and Huwaida T. I. Elshoush (2017) also proposed new chaos steganography and cryptography for audio files. They applied an (LSB) layers method for the plain message to encrypt it by a one-time pad algorithm. They used two chaotic sequences of Piecewise Linear Chaotic Map (PWLCM), one for generating the encryption key while the other to generate a random sequence for steganography process that used to embed the encrypted message in randomly selected audio samples [3]. While Jibran Hashim et.al. (2018) invested the AES encryption method as a backup in case the steganography algorithm has been broken, They used Least Significant Bit (LSB) modification technique as randomly bit spreading wise for secret message bits on different bits of the sample [9].

Some researchers tended to study encryption and try to improve encryption algorithms to get more immune systems, B. Ravi Kumar and Dr.P.R.K.Murti(2011) produced a Bit Shifting and Stuffing (BSS) methodology for encryption the text messages , The printable character needs 7 bits and the last bit value is 0 which is wasted in the character. In the BSS method, they are stuffing a new bit in the place of the wasted bit which is shifting from another printable character because the computer system requires one byte to represent a printable character i.e. 8 bits. After encryption, for every 8 bytes of plain characters, it will generate 7 bytes cipher characters and in decryption, Every 7 bytes of cipher characters will reproduce 8 bytes of plain characters[1].

3. Proposed System

The proposed system is a hybrid system aims to secure text messages through two security schemes: Encryption and Steganography. The text is encrypted by the Hill Encryption algorithm then hidden within a sound file in the time domain by changing the sign of the cover file values. The system consists of two phases illustrated in Figure(1):



Figure(1) The Proposed System

3.1. Cryptography Phase

The security first step is to encrypt the plain message by using a Hill algorithm which is a symmetrical key algorithm that uses a matrix(n×n) as a key and multiplies it by the ASCII code of the characters of the plain text, the encryption is done by the sender as follows:

a) Choosing an appropriate key which is a square matrix of n×n of an integer number and its inverse should be integer too for decryption phase.
   ex: key=[1 5 3;2 11 8;4 24 21].
b) Let M_plain is a plain text message needed to be secured, converted to ASCII code :
   M_plain='God pleas u!'
   asci_M=[71 111 100 32 112 108 101 97 115 32 117 33].
c) Reshape the asci_M to (m×n) Matrix.

A_M= [71 111 100; 32 112 108 ; 101 97 115 ; 32 117 33].

Note: if the the A_M can't be m×n dimension, It should be completed with zero's to make the multiplication done.

d) Multiply the asci_M with Key matrix to get Cipher_text by transforming ASCII to characters:
A=mod(A_M * Key,128).
A =[30 20 35 76; 115 112 13 79 ; 56 92 27 45].
Cipher_text=[-#Lsp←O8\-]

e) Hide the Cipher_text in a wav file and send it to the receiver.

f) end

At the receiver side, He extracts (unhide) the Cipher_text from cover file and decrypts it by reversing encryption steps means, multiplying the asci_code of Cipher_text by the inverse matrix of the encryption Key to get ASCII code of plain text and transforms it to characters then reconstructs a plain message again.

3.2. Steganography

The second step of the system is hiding Cipher_text in a cover file that is a wav file of two channels (stereo), using a substitutional method for hiding the Cipher_text in random positions in the second channel of the cover wav file, the steps of hiding algorithm is done as follows:

a. Taking the Ascii code of Cipher text and transform it to binary code :
[30 20 35 …]=[011110010100100011…]

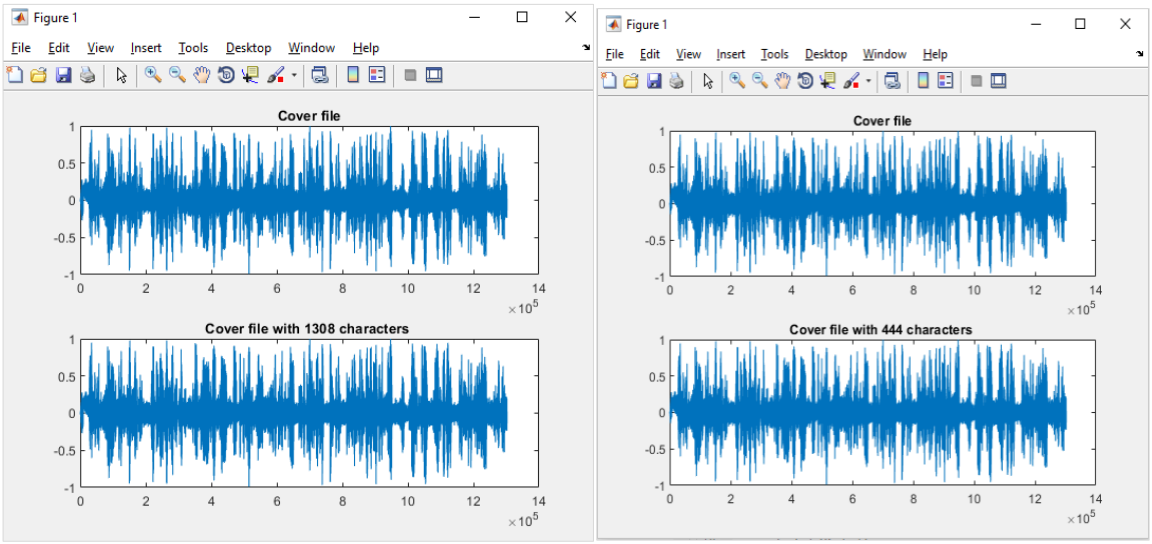
b. Calculating the number of bits of the cipher message and using a seed number (as a hiding key) exchanged between the sender and receiver to generate a number of random locations equal to the number of message bits to hide in.

c. The message bits are spread randomly in generated positions in such a way that :
$y(i) = \begin{cases} -|y(i)| & , bit(h) = 0 \\ |y(i)| & , bit(h) = 1 \end{cases}$, *y(i)* is the original sample of the cover file , i is randomly generated positions, *h =1,2,… No. of bits*.
$y = [\ 0.002 \quad -0.0019 \quad -0.0018 \quad 0.0016 \quad 0.0015 \ \ldots]$
$y = [-0.002 \quad -0.0019 \quad 0.0018 \quad 0.0016 \quad 0.0015 \quad \ldots]$

d. Reconstruct the wav file samples and send it to receiver.

e. End

The capacity of wav file to hide messages varies depending on sampling rate of the audio file, Each character need 7-bits to be represented so, If the sampling rate for a wav file is 48000 samples /second that means each second of wav file can carry over than 6000 characters, but in proposed system, the bits of the Cipher message are distributed randomly thus, they can't be noticed, hard to collect and arrange as well as the noise in sound file undetected, Figure(2) shows the wav file before and after hidden process.



(a) (b)

Figure(2) A cover wav file channel two that carry a cipher message with lengths: (a) has 1308 characters and(b) 444 characters

At a receiver side, the seed number is used to generate the locations of the message bits, collect them and form the ASCII code of the cipher text then decrypt it as in algorithm 3.1.

4. Results Analysis

The cover file in this paper restricted to be an audio (.wav) file format and the secret message is a text file. The system was tested for many sizes of audio files and message files, each time the size of cover file should be at least 7 times of text message file for ensuring the reliability of the system. The simulation is implemented with MATLAB R2018a software with core™ i7-3520M CPU@2.90GHz.
The measures PSNR, MSE, SSIM are calculated between cover file before and after embedding to ensure the imperceptibility requirement the results are shown in Table (1).

Table(1) PSNR,MSE,SSIM between the same cover before and after embedding

Size of message(character)	PSNR	MSE	SSIM
444	75.782	3.6657e-06	1.0
1308	61.4373	7.0823e-05	1.0
2170	60.67	9.8428e-05	1.0
2500	89.6534	9.9987e-09	1.0

As in Table(1) The system approved its efficiency in securing data and hiding it with least noise in the cover file.

The two important keys of the system are the encryption key that responsible for message distortion and the seed number that used for generating a series of random numbers to create locations for spreading message bits, To expose the message by the receiver he should know these two keys, otherwise, the mission is impossible, The sender and receiver exchange the keys between them either by dealing before or sending it secretly.

The time consumption for steganography and cryptography vary proportionally to messages length, and the time for decryption is less than that consumed in encryption for the same length of message, On the other hand the time needed for hide the same message is longer than its encryption time, as shown in Table (2)

Table(2) The elapsed time(with second) of security system vs message size

Size of message(character)	Encryption time	Decryption time	Hiding	Unhide
444	0.011	0.0128	0.027	0.069
1308	0.0208	0.066	0.091	0.0822
2170	0.0273	0.0715	0.159	0.106
2500	0.095	0.099	0.18	0.12

5. Conclusion

The security of the data is the most important issue in the digital world; Both Steganography and cryptography are together forming an integral system for security purposes. In our proposed system, the audio file is used as a cover file for securing an encrypted message. A Hill encryption method was used to encrypt the message, while audio steganography uses a new technique for hiding the message bits without changing cover values, It takes the signal of the audio samples (negative or positive) to act like 0 or 1 and randomly spreading the bits through a position generator seed as a hiding key. Both the keys must be private between the sender and receiver. This method is appropriate for securing any type of data like text or image with a minimum rate of distortion to the original sound wav and completely retrieving for the hidden messages and it was good in minimizing the noise in the cover file as measured by PSNR, MSE and SSIM for different lengths for tested messages with less time consumption.

References

[1] B. Ravi Kumar, Dr.P.R.K.Murti, (2011)," Data Encryption and Decryption process Using Bit Shifting and Stuffing (BSS) Methodology", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 7 ,pp: 2818:2827

[2] Huda adel ali , Enas wahab Abood and Wafaa A. Khudhair ,(2019) ,"Using LSB Method For Hiding Hill Encrypted Grayscale And RGB Images In RGB Image" , Journal of University of Garmian, Special Issue Conference paper SCPAS-2019,pp:46-53 .

[3] Samah M. H. Alwahbani and Huwaida T. I. Elshoush,(2017), "Hybrid Audio Steganography and Cryptography Method Based on High Least Significant Bit (LSB) Layers and One-Time Pad—A Novel Approach" Proceedings of SAI Intelligent Systems Conference, Intelligent Systems and Applications, pp: 431-453.

[4] K.Sakthisudhan, P.Prabhu and P.Thangaraj , (2012) ," Secure Audio Steganography for Hiding Secret information", International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012) Proceedings published in International Journal of Computer Applications® (IJCA),pp:33-37.

[5] Mustafa Sabah Taha, Mohd Shafry Mohd Rahim, Sameer abdulsattar lafta, Mohammed Mahdi Hashim and Hassanain Mahdi Alzuabidi,(2019), "Combination of Steganography and Cryptography: A short Survey", 2nd International Conference on Sustainable Engineering Techniques (ICSET 2019),IOP Conf. Series: Materials Science and Engineering 518 (2019) 052003,IOP Publishing ,doi:10.1088/1757-899X/518/5/052003, pp:1-13.

[6] Hashim M, Rahim M, Shafry M and Alwan A A ,(2018)," A review and open issues of multifarious image steganography techniques in spatial domain", Journal of Theoretical & Applied Information Technology 96.

[7] Chua Teck Jian, Chuah Chai Wen, Nurul Hidayah Binti Ab. Rahman , Isredza Rahmi and Binti A. Hamid,(2017) , "Audio Steganography with Embedded Text" International Research and Innovation Summit (IRIS2017) IOP Publishing , IOP Conf. Series, Materials Science and Engineering 226 (2017) 012084 doi:10.1088/1757-899X/226/1/012084.

[8] Mohammed Pooyan,Ahmed Delforouzi,(2007),"LSB based steganography method based on lifting WaveletTransform",IEEE international symposium on signal processing and information technology .

[9] Jibran Hashim , Arsalan Hameed , Muhammad Jamshed Abbas ,Muhammad Awais , Hassaan Aziz Qazi and Sohail Abbas, (2018),"LSB Modification based Audio Steganography using Advanced Encryption Standard (AES-256) Technique",12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS),IEEE, DOI: 10.1109/MACS.2018.8628458.

[10] Rina Mishra and Praveen Bhanodiya,(2015)," A Review On Steganography And Cryptography ", International Conference on Advances in Computer Engineering and Applications, IEEE, DOI: 10.1109/ICACEA.2015.7164679.

[11] Pooja P. Balgurgi and Sonal K. Jagtap,(2013),” Audio Steganography Used for Secure Data Transmission”, Aswatha Kumar M. et al. (Eds.): Proceedings of ICAdC, AISC 174 ,Springer India, pp. 699–706.
[12] K. Thangadurai ; G. Sudha Devi (2014),” An analysis of LSB based image steganography techniques” ,International Conference on Computer Communication and Informatics,IEEE Xplore, Coimbatore, India, DOI: 10.1109/ICCCI.2014.6921751.
[13] Krishna Bhowal · Debnath Bhattacharyya · Anindya Jyoti Pal · Tai-Hoon Kim,(2013) ,”A GA based audio steganography with enhanced security”,Springer Science+Business Media,LLC, Telecommun Syst (2013) 52 pp:2197–2204, DOI 10.1007/s11235-011-9542-0.
[14] S.S. Divya and M. Ram Mohan Reddy ,(2012),”Hiding Text In Audio Using Multiple Lsb Steganography And Provide Security Using Cryptography”, International Journal Of Scientific & Technology Research Vol. 1, Issue (6), ISSN 2277-8616.
[15] S. Hemalatha, U. Dinesh Acharya, A. Renuka, S. Deepthi and K. Jyothi Upadhya,(2015),” Audio steganography in discrete wavelet transform domain”, Journal of International Journal of Applied Engineering Research,Vol.10, Issue (16), pp: 37544-37549.
[16] Mazhar Tayel , Ahmed Gamal and Hamed Shawky, (2016),” A proposed implementation method of an audio steganography technique”, 2016 18th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, South Korea ,IEEE Xplore INSPEC Accession Number: 15824048 ,DOI:10.1109/ICACT.2016.7423320 .

Performance analysis of Google's Quick UDP Internet Connection Protocol under Software Simulator

Saif Talib Albasrawi

Education Collage /Misan University, Iraq
saiftalib@uomisan.edu.iq

Abstract: The QUIC protocol (Quick UDP Internet Connection) is a transport based over UDP Protocol (User Datagram Protocol) that provides safe, reliable and fast service on the Internet. Google proposed it to solve the problem of network delay. It is efficient, fast, and takes up fewer resources. The QUIC gathers the advantages of both TCP and UDP. QUIC is a user-level protocol running on top of UDP, which considered by IETF (Internet Engineering Task Force). Google assumes the response time of page load was short so that end-user performance was better. In this paper, the results presented on a local test using NS-3 (network simulator version 3) that allows QUIC output to test and design choices and possible limitations present the description of the functionalities and the main assumptions of the internet QUIC.
KEYWORDS: TCP protocol, UDP protocol, QUIC protocol, NS-3.

Introduction

Internet applications require reliable, fast communication and security. The TCP (Transmission Control Protocol) also provides the reliability and protection achieved by the TLS (Transport Layer Security) protocol. These protocols work in clients (user node) and servers (end nodes). Although the network infrastructure retains the high speed, it has to conserve for the applications. TCP and TLS have a critical role in user-comprehending performance [1].

QUIC is a network transport protocol; Google suggested this as a user-level protocol running over UDP instead of TCP, thus removing the necessity for the TCP protocol's initial handshake function. It runs its encryption scheme, which is comparable to TLS, combines link establishment and key agreement into 1 RTT only. In any case, QUIC can start a connection in 0 RTT, straight away sending encrypted application information or data to the server, when it already has in its cache the server declaration (from the previous connection). In any case, if QUIC already has the server declaration in its cache (from the previous connection), it can start a link in 0 RTT, send immediately encrypted application information or data to the server. Few research papers typically focused on the QUIC as transport for HTTP [2].QUIC only used by its services on Google's servers (G mail, Search, YouTube, etc.), and only Chrome and Chromium browsers initially support the QUIC protocol [3]. Lastly, NS-3 implementations and existing execution of congestion control algorithms rely on the configuration of the NS-3 TCP socket, which cannot reuse without the native implementation. In this manner, in the spirit of expressing the responsibility for implementing and updating the protocol, QUIC implementations for NS-3 would support the research community in networking due to their usability. Also, the QUIC stack intended to reuse the various TCP congestion management implementations within QUIC. To compare different congestion management designs for QUIC directly conceivable, while actual implementations give the New Reno adjustments only [4].In the remainder of this paper, an outline of the significant highlights of QUIC provided in Sec.2. Then, a description of the NS-3 implementation in Section 3, with more information on the code structure, QUIC compatibility with TCP and missing QUIC Internet-Drafts components, in Sec. 4, presents examples of QUIC performance evaluation. At last, the results concluded in Sec. 5.

Criteria of QUIC Transport Protocol

QUIC is a protocol for transport layers that aims to increase performance. It based on UDP that allows fast deployment of QUIC changes. Figure 1 shows how an application differs between TCP and QUIC [2].

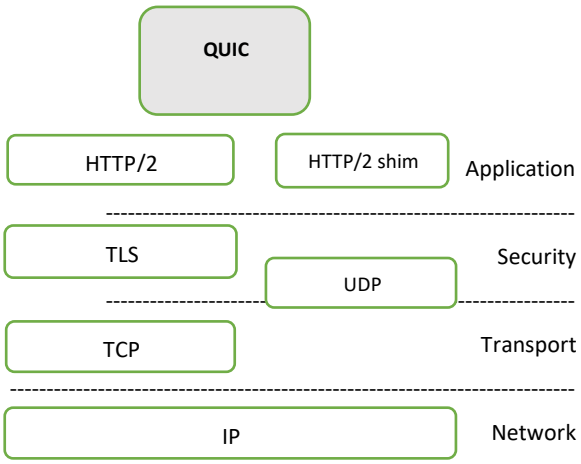


Figure 1. QUIC in the traditional stack

Two major design decisions make it possible to improve efficiency compared to TCP, QUIC integrates cryptography and handshake transportation to minimize setup latency and by default to provide a secure channel. To do this, it provides three types of connection setup [5, 6].

• **Initial handshake (Initial handshake 1-RTT):** Firstly, the client have no any server details. The client begins with a message about Client Hello (CHLO), which will discard with a REJ message by the server. It requires a server configuration with a long-term server authentication chain and a time stamp. Presently, the customer can send a complete message to CHLO containing the initial labels (tags) and the REJ messages. If the handshake is efficient, the server will be responding to an encrypted Hello message from the server (SHLO). The hello message for the server contains the ephemeral public value for the server that used to measure the ephemeral session key.

• **Repeat handshake (0-RTT):** In any past connection establishment, the client has only observed the REJ post. Which stored the labels (tags) from a REJ message in order to complete the Client Hello message (CHLO). Again, if the handshake is successful, the server will respond to an encrypted SHLO response. Through using the initial shared key, the two meetings will calculate the ephemeral keys to send and receive further messages. From that point, when the client wishes to reach 0-RTT latency, the request must encrypt with primary keys and send before receiving a server response. The server also stores the client's nonce and its public interest, which determines the shared key, to accomplish this.

•**Rejected 0-RTT (Failed 0-RTT)**: The server reacts with the REJ message when the information on the server has expired in complete CHLO. In this case, the 0-RTT try failed and the handshake proceeds as though it were an initial handshake.

The second performance-enhancing structure decision addresses the head of a line blocking issue. It occurs when a packet is lost during transit and has to retransmit. TCP's efficient service guarantees to have the packets delivered in the same order when they have sent. It makes TCP traffic vulnerable to header blocking; on delivery of the missing packet, subsequent packets must withstand. For every problem, QUIC uses different sources of lightweight structured abstractions. Each stream cut into frames. There are multiple forms of frames. Start with the standard stream frames contain data used to create connections (e.g., CHLO, REJ). It uses fixed stream 1. First, acknowledgment frames, which will address later, inform the sender of the successful delivery of packets. There is a congestion management unit, which is not included in our environment since the local network has only one single user. Ultimately, two types of frames are required to close this connection; they will be addressed later in section 4 [5].

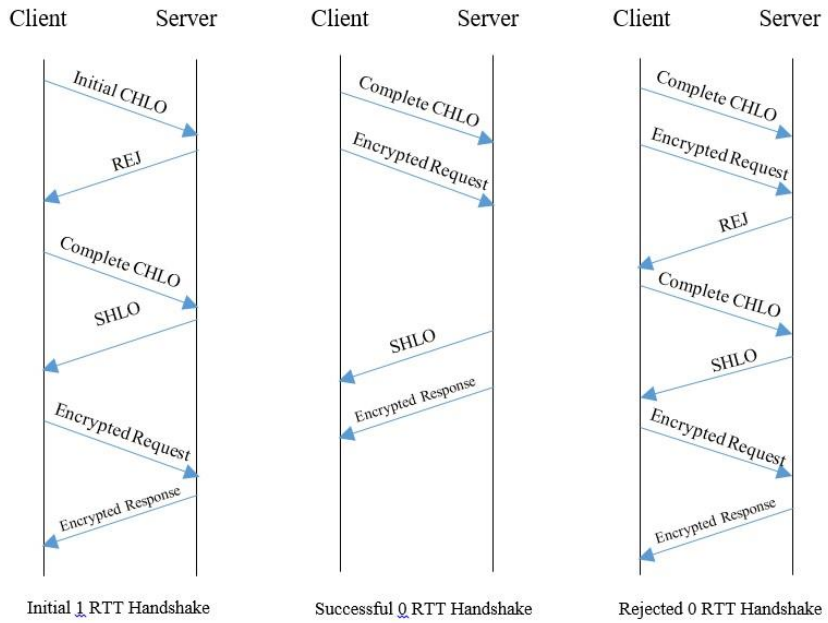


Figure 2. Outline of Various QUIC handshakes [6]

When a packet gets lost, it only affects certain streams that hold data in this packet. Subsequent data obtained for reassembly and delivery to the client.

2.1 important features of QUIC protocol include:

- *Communication Implementation Latency:* QUIC protocol integrates cryptography and handshake transport to reduce the amount of round trips needed for communication. It offers a cached client code that has reused to communicate with a previously considered server. It reduces the necessity for a new handshake.
- *Multiplexing:* QUIC multicast consists of streams, which transport data independently. Data for each stream in the specified frame sent using the stream identifier ID. A QUIC packet can consist of one or more frames.
- *Forward Error Correction:* QUIC supports FEC, where an FEC packet will contain the parity of the packet forming the FEC group. This function enabled or disabled if necessary. It allows it to recover the contents of a lost packet in an FEC group.
- *Connection Migration:* QUIC connection defined by the 64-bit connection identifier instead of 4 tuples from the IP addresses of source and destination and the essential communication port numbers. Thus, for example, if a device changes the Internet connection, if the IP addresses or NAT (Network Address Translation) connections change, the QUIC connection reused. QUIC allows encryption authentication of a relaying client, so using the session key, the client will encrypt and decrypt packets. [7, 8].

3. QUIC simulation

This section explains a simulation of the software. The benefit of QUIC protocol, and how the main classes of Quic protocol (internet module) can modify in the NS-3 to match UDP and TCP. Duplicating the isolation of congestion control from the primary socket and the presence of stand-alone buffer classes [9], while adding new components to TCP with representation for the novelties. As in the TCP execution, each client's Quic sockets models will instance a single Quic SocketBase entity, while the server will branch a new socket for all incoming connections. A Quic SocketBase object receives and transmits acknowledgments and packets, accounts for retransmissions, performs connection-level flow and congestion control, concerns the initial handshake, exchanges transport parameters, and manages the state machine and the life cycle of a QUIC link. The class Quic SocketBase provides pointers to many other related elements including:

A) Quic SocketTxBuffer and Quic SocketRxBuffer introduced socket transfer and reception buffers, respectively.

B) Quic SocketState object expanding TCP Socket State [9] with more variables used by the state and congestion control machines.
An Object extends the TCP CongestionOps class, which incorporates congestion control and makes it compliant with TCP's congestion control implementation. The QUIC socket connected to the main UDP socket via the Quic L4Protocol object, which also manages the initial development of a Quic SocketBase object, triggers the joining UDP socket, and the distribution of packets between the Quic SocketBase and UDP sockets. Instead a stream shown by the Quic StreamBase class which extends the basic Quic Stream class. It buffers application information, implements flow control at stream-level, and transmits the data received to the application. In addition, the Quic StreamBase object has pointers to transmit and receive the buffers to the full socket, which performed in the classes Quic StreamTxBuffer and Quic StreamRxBuff.
Various Quic StreamBase Objects connected via an entity to a single Quic SocketBase belong to the Quic L5Protocol class. The Quic SocketBase includes a pointer to an object Quic L5Protocol, and the last one carries a reference vector to different instances of Quic StreamBase. The Quic L5Protocol class establishes and configures the flows and issues of transmitting packets through flows and sockets to transmit and receive.

3.1 Packets, Frames and Headers

A QUIC packet consists of a header and a payload, as shown by projects on the QUIC Internet. The QUIC packet encapsulated in a wire-transmitted UDP datagram.
A packet header provided by the class Quic Header that extends the class ns-3 Header. When connected to the ns-3 packet, it deals with serialization and deserialization of the header and represents data transmitted by a QUIC header. One of the key novelties concerning the QUIC header structure's conventional TCP headers is that it can have two separate formats (long or short), as shown by the data calculation that should be shared. The passage in the packet number header can also have a variable length for short headers (one, two, or four bytes), depending on the number of bytes required to represent the packet number. For long headers (17 bytes), used to exchange messages that are essential to the link's lifecycle (such as initial handshake and version negotiation). Short headers (2-13 bytes) used in data packets and acknowledgments. The Quic Header class offers different static methods that can use to make various header types (long, short, for handshake message) and to set significant parameters. Alternatively, the Quic packet payload consists of several frames, every frame mapped to a stream and / or a control process for each frame also conveys an extra subheader enforced by the Quic Sub-header class, and indicates the frame type. Data frames properly connected to flow via a flow identifier, and their subheaders can convey frame size. Instead, the control frames used to perform specific actions; the ACK frames are the most control frames in the NS-3 Quic code; implement them as shown in [10,11], And it used to recognize packets that received at the endpoint of the link. Each ACK assigns the most significant number of accepted sequences and, when finding gaps in the received packet list, can hold multiple selective accepting blocks. In the ACK, the interval with the most important sequence number between the accepted packet receptions can list with the time it sent. The QUIC-protocol packet structure described in Figure 3 [12].

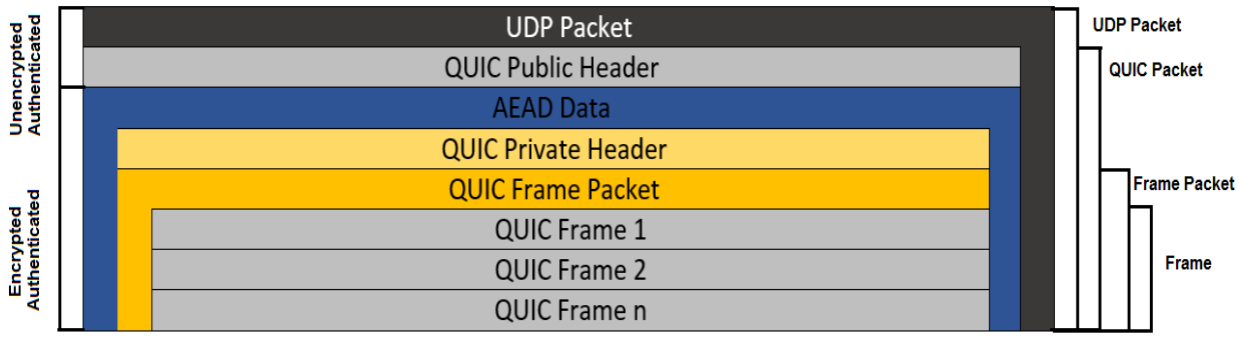


Figure 3. Packet structure of QUIC protocol

3.2 QUIC Buffers

- *In socket level*, The Quic SocketTxBuffer transmission buffer stores flow or control frames which Quic SocketBase passed into the socket and returned packets of the desired size. It comprises two separate types of packets, the first type to which it is still not transmitted, and the second type to which it is transmitted (until it is recognized) if retransmission is required. Each element stored in a Quic SocketTxItem list relates the intelligent pointer to the timing of the ACK receipt, the packet with sequence number and some flags that mark the packet when lost [13]. Additionally, it can separate and assemble frames depending on the specific packet size the socket requires. In the first transmission, when a packet is a transfer to the socket, the relevant frames removed from the unsent list; the entire packet has entered the sent list. In this case, it can control ACK receipt correctly, release packets exactly received, measure the number of bytes and handle retransmissions [14].
- *In stream level*, Quic StreamTxBuffer manages the sender side buffering. The transmission buffers store application packets until they accept it, but it does not retransmit, which based on the QUIC project's internet-draft, which stored all QUIC packets at the socket stage. In the stream, the packets buffer from the start of the stream marked by an offset. Since the received byte is already out of order, it will deposit to the stream buffer before the packet containing the missing byte received. As the stream buffer gets multiple bytes in an incorrect order, it passes them to the Quic SocketRxBuffer; it stores the received data packets released before the client requests them.

The stream includes the Finish bit (meaning the transmission has ended), which will recover the entire amount of bytes obtained in the stream level buffer. The concept that buffers in either Quic StreamBase or Quic SocketBase are always initialized [12, 15]. The buffers in this paper follow the TCP implementation method in ns-3, set the buffer size, the number of bytes that can manage buffering before rejecting a packet by using the socket features of those classes, and the stream buffers.

3.3 Retransmission Process

The QUIC retransmission mechanism works at the socket-level. It retransmits complete packets consisting of one or many frames that have lost by increasing the reception flow and ACK transmission used in QUIC implementation. In this case, ACK reception managed by the Quic SocketBase class during the RecAckFrame process and Quic L5Protocol begins when an ACK frame is in the received packet. The ACK frame periodically defines the most significant recognized number of packets, probably includes a few additional ACK blocks, which indicate missed packets [16]. The socket must pass the information to the Quic SocketTx buffer associated with it, which will mark it as missing or recognized, and restore the list to the socket. The socket performs proper congestion control activities at this stage by upgrading the socket status in case of loss or through the congestion window and activating the retransmission process, retransmissions in the QUIC protocol connected to the new one, monotonically-sequence numbers. Therefore, Quic SocketBase increases sequence number value by one each retransmission and advises the socket buffer. The last transfers the sent packets identified as missing back to the buffer that has not yet transmitted, updating their packet number. Finally, Quic SocketBase retransmits and forwards the packets to Quic L4Protocol [17].

3.4 QUIC compatibility with TCP congestion control

The modularity algorithm for congestion control according to the primary socket code mentioned in [9], This is one of the essential features of NS-3. It accomplished by using a specific object, which expands the congestion control functionality of a class with the fundamental TCP CongestionOps. Congestion management algorithms have introduced and joined to the Internet node. Vegas, for example, of TCP's NewReno. Since the legacy mode can use in Quic SocketBase. To push the congestion window for the QUIC link it can use TCP algorithms to manage congestion. It can use specific algorithms that have used additional details. Supported only for QUIC by the socket QUIC and the development of congestion management techniques. [18]. Compatibility with legacy TCP congestion control algorithms achieved by keeping the instance of TCP CongestionOps as the primary object of congestion control in Quic SocketBase. Later introduced the new Quic CongestionControl class, which extends TCP NewReno. Quic CongestionControl has other methods which include additional information that QUIC Internet-

Draft defines in the congestion control algorithm (for example, packet transmission, more accurate RTT information) if required[19].

3.5 Connection Structure

The connection of QUIC is one communication between two endpoints, each defined by a combination of IP addresses and UDP ports in a specific way [11]. The QUIC system introduced an interactive internet-draft with open connections and introduced client-server version matching and handshaking. In addition, the procedure implemented models handshake but did not explicitly encrypt. The QUIC will handshake based on the combined endpoint date in two separate ways. Upon recently authentication of the endpoints, 0-RTT handshaking can use. It models by attaching a list of checked endpoints to Quic L4Protocol that tested before the first packet sent and at another endpoint when it received. In addition, 0-RTT was added in Quic L4Protocol using the Handshake attribute 0-RTT. If the 0-RTT not enabled, then the socket reverts with the original client message to a 1-RTT handshake. Lastly, if the two endpoints in a version differ, a 2-RTT handshake will use to stabilize the point-to-point release of QUIC [20, 21].

3.6 Applications and Testing

Finally, an appropriate set of unit tests performed; it will expand in future releases. The test suite (quic-tx-buffer) first tests the correctness of the buffering behavior of stream and socket transmission, implemented by Quic StreamTxBuffer and Quic SocketTxBuffer. The test checks various inserts for both stream and socket and removes events from buffers, including tests where the insert failed due to the presence of the entire output buffer and also, the packet must reinsert in the stream buffer. In addition, socket checking in case of retransmission verifies the behavior. Likewise, the second test community (Quic Rx Buffer) performs on the socket buffers to add and remove operations and stream reception. The third set of tests (the quic header) finally checks Quic Header and Quic Subheader implementations. In addition, to directly use the streaming multiplexing features, which introduced the adaptation of Quic protocol's UDP client-server (Client and Server) applications, allowing the API to send to the socket using the stream identifier. So, the data scheduled through the available streams.

4 Test Results

Now, it presents a set of preliminary results for testing NS-3 QUIC implementation using various congestion control algorithms in both the legacy and non-legacy modes. By using standard dumb-bell topology, also used to test TCP performance [22] when a common 5 Mbps bottleneck is used between clients and servers. In our example, the minimum RTT is 50 milliseconds used to implement this experiment has been a quic-variants-comparison, which has changed from TCP. For two old congestion management algorithms (i.e., New Reno and Vegas), with the non-legacy alternative New Reno adapts the congestion avoidance stage to a slightly different congestion window. The development of congestion windows corresponds to the predicted behavior of the BDP product (Bandwidth-Delay Product) algorithms with two steady-state flows (20 kB). Furthermore, when comparing the actions associated with delay, congestion (i.e., Vegas) and loss (i.e., New Reno and QUIC), you can see that Vegas has managed to save less RTT, as planned. Lastly, congestion management of New Reno and QUIC showed similar patterns. The latter distinguished at the congestion avoidance level by improved window width.

1.1 Comparison of performance between TCP and QUIC

This test performed to compare the effect of various network connection parameters on TCP and QUIC transmission. The purpose of this test is to examine the behavior of QUIC, which is the same or better than TCP, by comparing protocol interactions with different network bandwidth. Then to run the same test, but with TCP, the bytes sent had the same file size as used in the QUIC test. Traffic shaping also performed using the same test in the QUIC protocol. The throughput test in Figure 4 and Figure 5 shows the different throughput of TCP and QUIC with various Bandwidth and RTT.

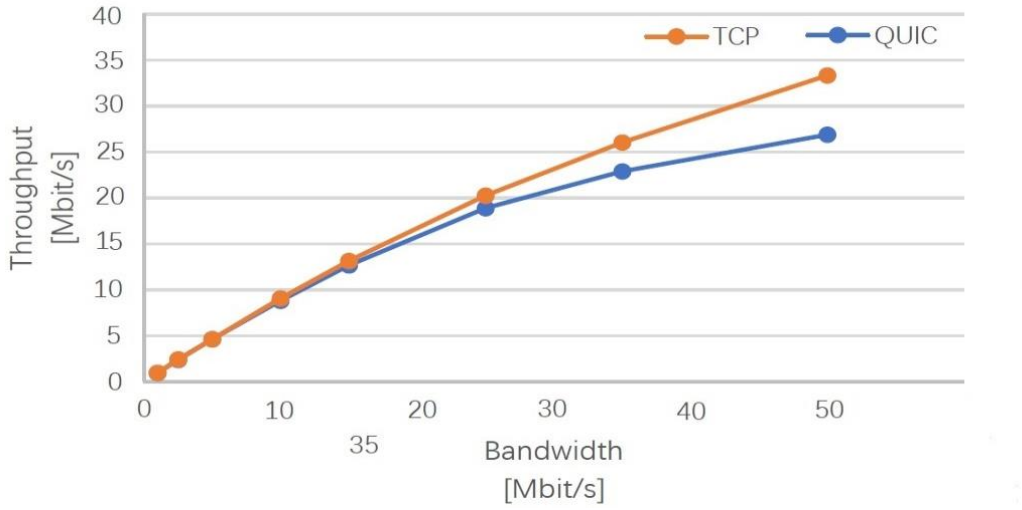


Figure 4. Comparison between TCP and QUIC

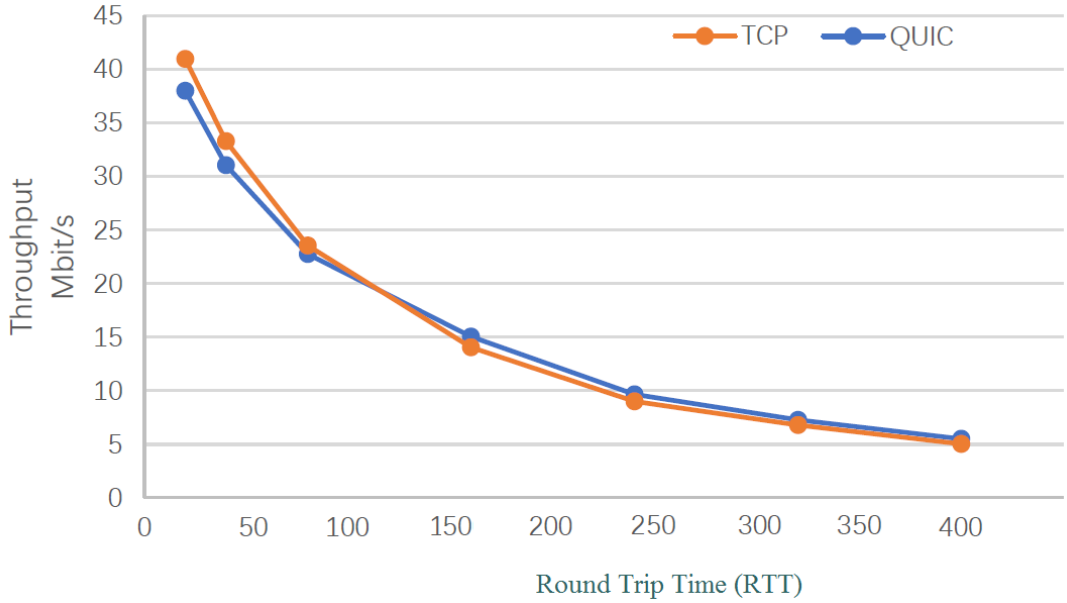


Figure 5. Test 0-400ms Round Trip Time (RTT), 50Mbps and 0%loss

5 Conclusions

In this paper, the original QUIC implementation introduced on the base of NS-3, which is the new IETF standardized transport protocol and represents the majority of Internet traffic. The QUIC application mentioned in this paper expands the structure of the TCP code in NS-3 with functions that differentiate QUIC, such as stream multiplexing, and initial handshakes with low latency. In addition, improve the QUIC socket framework so that both the latest QUIC-only congestion control algorithms and the legacy TCP can plugin. Testing the implementation using simulations, a dumb-bell topology showing the standard behavior of congestion control algorithms also compared their performance with a non-legacy internet.

References

[1] Rasool A, Alpar G, and de Ruiter J 2019 State machine inference of QUIC arXiv preprint arXiv:1903.04384.
[2] Carlucci G, De Cicco L and Mascolo S 2015 HTTP over UDP: An experimental investigation of QUIC in Proc. of the 30th Annual ACM Symp. on Applied Computing-SAC Vol. **15** pp 609–614
[3] Cook S, Mathieu B, Truong P and Hamchaoui I 2017 QUIC: Better for what and for whom? IEEE Int. Conf. on Communications (ICC) pp 1-6
[4] Kakhki A, Jero S, Choffnes D, Nita-Rotaru C and Mislove A, 2017 Taking a long look at QUIC: an approach for rigorous evaluation of rapidly evolving transport protocols in Proc. of Internet Measurement Conf. pp

290-303
[5] Lychev R, Jero S, Boldyreva A and Nita-Rotaru C 2015 How secure and quick is QUIC? Provable Security and performance analyses IEEE Symp. on Security and Privacy (San Jose: IEEE) pp 214-231
[6] Langley A, Riddoch A, Wilk A, Vicente A, Krasic C, Zhang D, Yang F 2017 The quic transport protocol: Design and internet-scale deployment in Proc. of the Conf. of the ACM Special Interest Group on Data Communication pp 183-196
[7] Das S 2014 Evaluation of QUIC on web page performance Doctoral dissertation, Massachusetts Institute of Technology
[8] Stone C, Chothia T and de Ruiter J 2018 Extending automated protocol state learning for the handshake. In European Symp. on Research in Computer Security (Springer Cham) pp 325-345
[9] Casoni M and Patriciello N 2016 Next-generation TCP for ns-3 simulator *Simulation Modelling Practice and Theory* vol **66** pp 81-93
[10] Iyengar J and Swett I 2018 Quic loss detection and congestion control Internet Engineering Task Force.[Online]. Available from: https://datatracker. ietf. org/doc/html/draft-ietf-quic-recovery-07 [Accessed 19 Mar 2020]
[11] Iyengar J and Thomson M 2018 Quic: A udp-based multiplexed and secure transport Internet Engineering Task Force. [Online]. Available from: https://tools.ietf.org/html/draft-ietf-quic-transport-17 [Accessed 19 Mar 2020]
[12]Bishop S, Fairbairn M, Norrish M, Sewell P, Smith M and Wansbrough K 2005 Rigorous specification and conformance testing techniques for network protocols, as applied to TCP, UDP, and sockets in Proc. of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications vol. **35** pp 265–276
[13] Ekiz, N, Rahman A and Amer P, 2011 Misbehaviors in TCP SACK generation ACM SIGCOMM Computer Communication Review, vol. **41**(2) pp16-23
[14] Thomson M and Turner S 2019 Using TLS to Secure QUIC [Online]. Available from: https://quicwg.org/base-drafts/draft-ietf-quic-tls.html [Accessed 19 Mar 2020]
[15] Shade R and Warres M 2016 HTTP/2 Semantics Using The QUIC Transport Protocol IETF. [Online]. Available: https://tools.ietf.org/html/draft-shade-quic-http2-mapping-00 [Accessed 19 Mar 2020]
[16] Megyesi P, Kramer Z and Molnar S 2016 How quick is QUIC? in IEEE Int. Conf. on Communications pp. 1-6
[17]Carlucci G, De Cicco L and Mascolo S 2015 HTTP over UDP: an Experimental Investigation of QUIC in Proc. of the 30th Annual ACM Symp. on Applied Computing pp 609-614
[18] Nguyen T, Gangadhar S, Rahman M and Sterbenz 2016 An Implementation of Scalable, Vegas, Veno, and YeAH Congestion Control Algorithms in ns-3 in Proc. of the Workshop on ns-3 pp 17-24
[19] Hamilton R, Iyengar J, Swett I and Wilk A 2016 Quic: A UDP-based secure and reliable transport for http/2.IETF. [Online]. Available from: https://tools.ietf.org/id/draft-tsvwg-quic-protocol-02.html [Accessed 19 Mar 2020]
[20] Fischlin M and Gunther F 2014 Multi-stage key exchange and the case of Google's QUIC protocol in Proc. of the ACM SIGSAC Conf. on Computer and Communications Security pp 1193-1204
[21] Ruth J, Poese I, Dietze C and Hohlfeld O 2018 A First Look at QUIC in the Wild in Int. Conf. on Passive and Active Network Measurement (Springer Cham) pp 255-268
[22] Fiterau P, Janssen R and Vaandrager F 2016 Combining model learning and model checking to analyze TCP implementations in International Conference on Computer-Aided Verification (Berlin: Springer Cham) pp 454-471

Measuring the Impact of Using Different Tools on Classification System Results

Zainab A. Khalaf[1] **Zainab M. Jawad**[2]

Basrah University, College of Science, Department of Mathematics E-mail: zainab.ali2004@gmail.com
Basrah University, College of Education for Pure Sciences, Department of Computer Science E-mail: z82m89asia@gmail.com

Abstract. A huge amount of textual data is available on the web. These data need to be classified under labels or classes to make the search more efficient and easier. Achieved by using automatic classification is used for this task. Many factors impact on the performance of the classifier system, such as the amount of using dataset, the data dispersion degree, preprocessing tools, feature extraction methods, terms weighting, and data reduction. So, researchers constantly compete to build a robust classifier with good performance. This study focuses on the effect of using different tools in preprocessing and term weighting stages. The experimental results applied on two different languages (Arabic and English languages). Also, the experimental results were compared with the recent related works.

1. Introduction

The amount of electronic textual data has rabidly increases, and this data is not organized based on categorizes. Consequently, searching and accessing the required data takes time. So, we require an automatic system capable of classifying data effectively and efficiently [1]. The text classification is the task that organizes the textual document (e.g. DOCX, HTML, TXT, PDF files) into classes based on predefined training documents. A lot of supervised machine learning was used as classifier like Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Trees, etc [2]. Many factors affects the classifier result, such as the amount of data training used, preprocessing tools (e.g. stemmer, part of speech tagging), feature extraction methods, and high dimensionality reduction factors. Generally, a large amount of data used to train the classifier for enhancing the classification result. However, this will cause a high dimensionality problem by increasing the number of irrelevant feature, which could negatively affect classification performance. Therefore, the feature reduction can be used to reduce the data dimensionality and improve the performance of the classification system [3].

The main objective of this paper to study the impact of tools on classification performance by using two feature reduction approaches. Also, the experimental results will compared with recent other research on the same datasets.

The organization of the paper is: the section (2) review the related work, in the section (3), the classification system will be explained, while the section (4) shows the proposed system method. The section (5) will discuss the experiments and the results, and then follows by the discussion of the results in section (6). Finally, the section (7) will summarize the conclusions and the future works.

2. Related Works

Researchers have improved the text classification either by reducing the features or enhancing the classification performance. Those tools could be the dataset, stemmers, number of stop words, feature reduction methods or the classification algorithm. The tools used in the classification will affect the classification positively or negatively; those tools could be the dataset, stemmers, number of stop words, feature reduction methods or the classification algorithm. Wan et al.

[4] used bigram (which extracts two features that appear continuously), 2-termset (the two terms that occur in the same documents considered one feature) and SABigram (tokenizing the documents based on the stop words and the punctuations) in the preprocessing stage and applied the NB and SVM classifier on the 20-newsgroup dataset, the F1-score result was 64% and 82% respectively. also Dogan et al [5] used the 20-newsgroup and they used the tokenization based on white space and Porter Stemmer then applied the SVM, KNN and NN the F1-score results was 91%, 64% and 82% respectively.

The features affect the classification performance extensively because of the high dimensionality caused by the large vocabulary, especially in the big datasets. Therefor, many researchers try to reduce the features using different methods. Wan et al. [4] proposed a feature selection method called relevance category frequency (rcf) that considers the strength of the composite features and the ability to distinguish. They then combine the rcf with the Chi-square (Chi2). Furthermore, the researchers used three methods for feature extraction bigram, 2-termsets and the SABigram . The main disadvantage of the SABigram is that it is hard to find the entire combined features depending on the stop words and the punctuation tokenization; also, the rcf must combine with other feature reduction methods. Chandra [6] used three feature reduction methods the Chi2, Mutual Information (MI) and Term Frequency (TF). They applied these on three datasets using SVM and NB classifiers. They considered Chi2 as the best accuracy than the others. While Rehman et al [7] proposed a new feature selection method named Normalized Difference Measure (NDM) compared results with seven feature selection methods and seven datasets using two classifiers the SVM and NB. However, the NDM shows lower performance in the large datasets compared with the other feature reduction methods. On the other hand, Yang et al [8] compared the performance of six feature reduction methods: Information Gain (IG), Chi2, MI, Odd Ratio (OR), Gini Index (GINI) and Expected Cross Entropy (ECE) with their proposed method. During the preprocessing step, they depended on finding the probability of the feature in each class. If the probability was greater than the mean value, then the feature was selected. They applied the feature selection methods on the features extracted from the preprocessing for comparison and improved the performance of the original feature selection methods. However, they only used the Na¨ıve Bayes classifier to prove their method; they should try other classifiers for experimental verification. While, Mowafy et al [9] used the Chi2, IG, OR, Distinguishing features selector (DFS) and GINI feature selection and compared the results with the TF-IDF weighting, using the Multinomial NB and KNN. The result shows the preference for using Chi2, but they did not compare this method with other feature selection methods. In (table 1) shows a related works with more details.

In this study we will compare the classification performance using TF-IDF and Chi2 methods.

Moreover, we will compare our results with other studies to explain the different implementation using different functions.

Table 1. The Related Works

Reference No.	Year	Tools For Preprocessing	Feature Extraction Methods	Classification Model	No. of Features	Datasets Used For Evaluation	No. Of Classes	F1-Score	Accuracy
[4]	2019	bigram 2-termset SABigram	Chi2 · rcf	SVM NB	5500 6705	20-newsgroup (19997) Ruters-21578	20 10	82% 64% 93% 56%	
[6]	2019		Chi2, MI, TF	SVM NB	20% 20%	20-newsgroup (18846) Reuters-21578	20 10	92% 92% 86% 87%	
[5]	2019	Tokenization Normalization stop list stemming (Porter Stemmer)	TF-IDF	SVM KNN NN	25000 8000	20-newsgroup (19997) Reuters-21578	20 10	91% 64% 82% 87% 75% 81%	
[8]	2019	Tokenization Normalization stop list stemming	IG, Chi2, MI, OR, GINI, ECE,	NB	2000 1500	20-newsgroup (19997) Reuters-21578	20 10	78% — 84%	
[10]	2019		TF more than 2	SVM NB	43553 50570	20-newsgroup (19997) 20-newsgroup (18828)	20 20		61% 76% 85% 86%
[11]	2018	Tokenization Normalization stop list stemming	TF-IDF	NB + KNN — SVM+uni-gram	4856 10913	20-newsgroup (19997) Reuters-21578	20 50		87% 86% 88%
[7]	2017		IG, Chi2, OR, DFS, GINI	SVM NB	1500	20-newsgroup (19997)	20	75% 73%	

3. Text Classiftcation System

The system that assign a group of textual documents into one or more predefined classes based on their subjects automatically called text classification system. Let D is set of learning documents that belong to different classes, where D =$\{d_1 , \ldots , d_n\}$ with predefined classes
C = $\{c_1, \ldots , c_m\}$, n and m are the number of documents and predefined classes respectively.
And, suppose x is a new document, so the classification system tries to predicate class (c_x) to document x based on the most likely similarity classes in C, that is c_x in C.
The text classification system consists of five main steps, as shown in figure (1). ***Data Collection*** is a set of unstructured documents collected from different resources used to learning the classifier model. ***Preprocessing*** is the primary step used to prepare the unstructured documents for subsequent processing. ***Representation and Dimension Reduction*** are the processes aiming to repesent the weighting features and reduce the high dimensionality of the data. ***Classification Algorithms*** are the technique used to predict the class label of the new document based on the predefined pattern, ***Performance Evaluation*** is a process used to measure classifier performance.

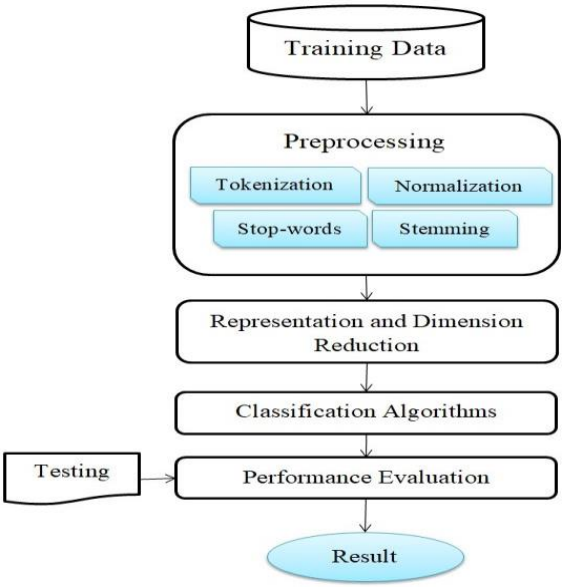


Figure 1. The Text Classification System

4. The Proposed System

The text classification is automatically labelling the new documents based on the training of predefined labelled documents. The first step of any classification system is preparing the input documents for processing. Then, the critical features are extracted and passed as input for the classifier. Finally, we evaluate the classifier results in the next stage, explained in the following subsections:

4.1. Preprocessing Stage

The preprocessing aims to prepare the data collection for processing [12][13]. Preprocessing includes many subtasks:
Tokenization: is cutting the documents into tokens by using the space for separation. *Normalization*: a subtask used to exclude the white space, special symbols and characters, as well as unify the shape of letters, such as convert uppercase letters to lowercase letters for English language, or unify the shape of some of the Arabic letters. *Stop-words*: the list of the words that frequently appear in the language like conjunctions and prepositions. These words are removed to reduce the space model of classification. *Stemming and lemmatization*: exclude suffixes and prefixes from words to reduce the word forms. The current study, uses Stanford corenlp lemmatizer for lemmatization. While, porter stemmer and ArabicLightStemme are used for English and Arabic, respectively. The remaining terms obtained from the preprocessing step called features are used to create the vector space model (VSM). The VSM dimension

is n*m, where n is the number of classes, and m is the number of features. Each cell in VSM represents the weight of the feature (f_j) in the class (c_i) [14][7].

4.2. Dimension Reduction

The performance of text classification depends on the amount of trained dataset. Moreover, a large dataset causes high dimensionality. Therefore, dimension reduction is needed to decrease the computations budget without impacting the classifier performance. For this reason, the important features can be extracted from the large dataset and thus exclude irrelevant features. TF-IDF and Chi2 consider one of the effective methods to weighting and extract features by selecting the best features weight.

(i) TF-IDF: is calculated by finding the TF and IDF as the following equation [15]:

$$TF = \log_2(tf_{i,j}) \quad (1)$$

tf_{ij} is the term frequency of term i in document j

$$IDF = \log_2\left(\frac{N}{n_j}\right) \quad (2)$$

N is the number of documents in the collection, n_j the number of documents containing the term j.

And the TF-IDF will calculated as the following:

$$TF - IDF = TF * IDF \quad (3)$$

(ii) Chi2: Chi-square is one of the main feature reduction methods. The basic idea of Chi2 is to compare the distribution of actual data and the distribution of expected data. Suppose C is the classes c_1, c_2, ..., c_i, and F is the feature f_1, f_2, ..., f_r. The data documents will described by C and F. Let (C_i, F_j) denote the event that feature C takes on value c_i and feature F takes on valuef_j. The equation will calculate Chi2 [16][17]:

$$Chi2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (4)$$

Where n: the expected frequency (actual), e: the frequency that gained (observed). The e_{ij} calculated by:

$$e_{ij} = \frac{(n_i n_j)}{n}$$

n_i: the number of the documents that belongs to the class i. n_j: the number of documents that have the feature j.

4.3. Classification Algorithms

The classifiers used in this study are Naïve Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN).

Naïve Bayes (NB): NB is a simple and efficient probabilistic algorithm based on the distribution probability between the documents and the classes. The training phase obtains a set of parameters, which are used in the testing phase to predict the probability of the document belonging to the class [12][18].

***Support Vector Machine (SVM)*:** SVM is a stable and effective statistical classifier. The main idea of the SVM is using the training data to find the hyperplane with the highest Margin, where the margin is the largest distance between the nearest points on both sides of the hyperplane. The hyperplane used in the testing phase to decides the class of the new document [19][20][21].

***K-Nearest Neighbors (KNN)*:** KNN is a simple and easy to understand algorithm. The main idea is to find the distance between the new document and the training data documents and vote for the k nearest (smallest distance) documents. There are different measures used like Euclidean distance, cosine measure, etc. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document [22].

4.4. Performance Evaluation

The main performance measures widely used in text classification are the Precision, Recall and F1-csore. Precision (P) is indicated as the ratio between the number of properly classified documents (Correct Classes) to all documents that can be automatically recognized by the class (Predicated Classes). The recall (R) is defined as the ratio between the number of properly classified documents (Correct Classes) to all documents belonging to that class (Actual Classes). The F1-score indicates the harmonic average between precision and recall [23][24].

$$Precision = \frac{CorrectClasses}{PredicatedClasses} \quad (6)$$

$$Recall = \frac{CorrectClasses}{ActualClasses} \quad (7)$$

$$F1\text{-}score = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \quad (8)$$

5. Experimental Results & Discussion

In this section, the performance of the proposed system will be evaluated. TF-IDF and Chi2 are used for feature selection; different percentages are used for dimension reduction by selecting the best features. Also, we compare many related works using the same datasets.

5.1. Datasets

Two languages are used in the current study: English and Arabic. For the English language, 20-newsgroups [25] and Reuters-21578 [26] are used; for the Arabic language, Watan-2004 [27] and Khalaf-2018 are used. The 20-newsgroup contains twenty classes with 19997 documents, and the top seven classes from Reuters-21578 data in 10806 documents are used. The Watan-2004 dataset contains around 20291 documents in six classes, while Khalaf-2018, which is collected manually by [28] from the BBC and CNN online websites containing of 2750 Arabic news belongs to six classes. All these datasets are divided into 75% documents for training and 25% for testing.

5.2. The Classification System Results

After applying the preprocessing stage used to create the VSM, the features extracted from this stage are 86151, 18431, 78130 and 31980 for 20-newsgroup, Reuters-21578, Watan-2004 and Khalaf-2018, respectively. Three classification systems (NB, SVM and KNN) are applied with the parameters for NB is the default multinomialNB, linear SVM and KNN using Euclidian distance when K=5.

The experiment examined six percentages of reduction (80%, 70%, 60%, 55%, 50%) from the original features by applying TF-IDF and Chi2. The best result was with the 55% reduction. The number of feature reduced to 47383, 10137, 42971 and 11193 for 20-newsgroup, Reuters-21578, Watan-2994 and Khalaf-2018, respectively. The (table 2) shows the results of the classification system. The F1-score was enhanced for NB and SVM classification system, The KNN performs the worst. The reason is because of the neighbor features may introduce more noisy information instead of useful information.

Table 2. The Evaluation Of Current Study

.	20-Newsgoup		. Reuters-21578		. Watan-2004		. Khalaf-2018	
No. Of Features Befor Reduction	86151		18431		78130		31980	
No. Of Features After Reduction	47383		10137		42971		11193	
	TF-IDF	Chi2	TF-IDF	Chi2	TF-IDF	Chi2	TF-IDF	Chi2
NB	91%	91%	88%	82%	88%	88%	73%	76%
SVM	95%	95%	93%	93%	94%	95%	91%	91%
KNN	80%	60%	83%	79%	93%	86%	91%	85%

6. Discussion

This study used three classifiers to classify four datasets, then applied two feature reduction methods: TF-IDF and Chi2. Feature selection with different rank ratios (50%-80%) are evaluated in this study is to find the best ratio for evaluation. The best rank ratio obtained using the TF-IDF and Chi2 is 55%. When comparing the TF-IDF and Chi2 feature reduction methods, the TF-IDF shows a better enhancement in the English datasets, while the Chi2 was the best for Arabic datasets.

In the other hand, the (table 3) shows other researchers' results on the using datasets. The F1-score results of the current study show a better performance using the four datasets over the results of the other studies. Also, the table indicates the difference in the preprocessing tools and how the stemmers and stop words effect the classification performance

For the 20-newsgroup dataset, Dogan et al [5] split the data into 50:50 for training and testing. Porter stemmer and TF-IDF reduction (the number of features after reduction was 25000) were used, and then the obtained features were passed to the SVM, KNN and Neural Networks (NN) classifiers. Despite their number of the features being less than the current study, the accuracy result of the current study is better. Furthermore, Wan et al [4] proposed

a feature selection method called Chi2 and rcf. The accuracy results of current study is better than thier results. As mentioned in section 2, they used three methods of feature extraction. However, some disadvantages could affect the classification performance. Also, the current study showed a better accuracy performance than the results in [4] and [5].

For Reuters-21578, the researchers Unnikrishnan et al [29] also used the top seven classes from Reuters-21578. In the preprocessing, they used a Porter Stemmer algorithm with removing the words that occur at least in all the classes but one. Furthermore, the SVM and KNN were applied to compare with their proposed system, while the current study applied the Stanford lemmatizer and Porter stemmer, and the entire feature extracted was used and achieved better performance.

In the Watan-2004 dataset, the researchers Sabbah et al [30] selected 9,000 document from the original dataset and applied the Arabic Light Stemmer with using Chi2, IG and proposed SVM-FRM for feature selection and got 93% for SVM classifier. The current study, used all the documents of the datasets and got a better result of 94% for SVM classifier.

For Khalaf-2018, the researcher Khalaf et al [28] used the Khoja stemmer from safar library and the stop list containing (10,471) words in the preprocessing stage. Their SVM result was 91% using 50% from the features. While the current study used the same dataset by applying Stanford core nlp and ArabicLightStemmer, the stop list contained (15,847) words in the preprocessing stage. The results of the current study were 91% with less number of features.

Table 3. Comparing The Results Of Our Study With Other Researchers Studies

Reference No.	Year	Tools For Preprocessing	Feature Extraction Methods	Classification Model	No. of Features	Datasets Used For Evaluation	No. Of Classes	F1-Score
[5]	2019	Tokenization Normalization stop list stemming (Porter Stemmer)	TF-IDF	SVM KNN	25000	20-newsgroup (19997)	20	91%, 64%
[29]	2019	Porter stemmer and remove the words that occur in all the classes	TF	SVM KNN	1500	Reuters-21578	7	86%, 85%, 78%
[30]	2011	Arabic Light stemmer	Chi1, IG SVM-FRM	SVM	5000	Watan-2004 (9000 documents)	6	93%
[28]	2018	khoja stemmer (10471) stop words.	N-gram	SVM	12474	Khalaf-2018	6	91%
Current study		Porter stemmer Stanford Lemmatizer remove the words that occur in all classes (797) stop words ArabicLightStemmer Stanford Lemmatizer remove the words that occur in all classes (15847) stop words	TF-IDF CHI2	NB SVM KNN	47383 10137 42971 11193	20-newsgroup Reuters-21578 Watan-2004 Khalaf-2018	20 7 6 6	91%, 95%, 80% 88%, 93%, 83% 88%, 94%, 93% 73%, 91%, 91%

7. Conclusions and Future works

The most important issue in the text classification is efficiency and the reliability. Therefore, the researchers who study text classification try to produce the tools and functions that perform efficiently. In this study, we compared the performance of two feature selection methods TF-IDF and Chi2 by finding the best percentage of reduction from the original feature number. The best F1-score result was in the 55% feature reduction in both the TF-IDF and Chi2. However, the English datasets show better TF-DF results than the

Chi2, unlike the Arabic dataset, which found that Chi2 was best. Also, we compared results with other recent research findings to illustrate how the tool affects classification performance. Thus, it is important to find the best tools when building a classification system Future research could improve the classification system using other feature reduction methods, different classifiers or by enhancing the functions used in the text classification system (e.g. stemmers, feature reduction or classifiers).

References

[1] Ge Lihao & Moh Teng-Sheng, (2018), "Improving Text Classification With Word Embedding," Proceedings - 2017 IEEE International Conference on Big Data, pp. 1796–1805.

[2] Bai Rujiang & Wang Xiaoyue & and Liao Junhua, (2010), "Extract Semantic Information From Wordnet To Improve", Advances in Computer Science and Iformation Technology conference, Springer-Verlag Berlin Heidelberg NewYork, Vol. 6059, pp. 409–420.

[3] Kadhim Ammar, (2019), "Survey On Supervised Machine Learning Techniques For Automatic Text Classification", Springer Artificial Intelligence Review, Vol. 52, pp. 273–292.

[4] Wan Chuan, Wang Yuling, Liu Yaoze, ji Jinchao & Feng, Guozhong, (2019), "Composite Feature Extraction And Selection For Text Classification", IEEE Access, Vol. 7, pp.1-1.

[5] Dogan Turgut & Uysal Alper Kursat, (2019), "Improved inverse gravity moment term weighting for text classification," Elsevier, Expert Systems with Applications, Vol. 130, pp. 45–59.

[6] Chandra Andreas, (2019), "Comparison Of Feature Selection For Imbalance Text Datasets," International Conference on Information Management and Technology, Vol. 1, No. August, pp. 68–72.

[7] Rehman Abdur, Javed Kashif & Babri Haroon, (2017), "Feature Selection Based On A Normalized Difference Measure For Text Classification", Elsevier Information Processing and Management, Vol.53, No. 2, pp. 473–489.

[8] Yang Jieming, Lu Yixin & Liu Zhiying, (2019), "An Improved Strategy of the Feature Selection Algorithm for the Text Categorization," International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, No. 2, pp. 3–7.

[9] Mowafy M, Rezk A & El-bakry HM, (2018), "iMedPub Journals An Efficient Classification Model for Unstructured Text Document Keywords : Introduction", American Journal of Computer Science and Information Technology, Vol. 6, No.1:16 pp.1-10.

[10] Kilimc Zeynepi & Akyokus Selim, (2018), "Deep Learning- And Word Embedding-Based Heterogeneous Classifier Ensembles For Text Classification", Wiley Hidawi Complexity, Vol. 2018, pp. 1–10.

[11] Panda Mrutyunjaya, (2018), "Developing an Efficient Text Pre-Processing Method with Sparse Generative Naive Bayes for Text Mining," International Journal of Modern Education and Computer Science, Vol. 10, No. 9, pp. 11–19.

[12] Kobayashi Vladimer & Mol Stefan & Berkers Hannah & Kismihók Gabor & Hartog Den, (2018), "Text Classification For Organizational Researchers: A Tutorial", Organizational Research Methods, Vol. 21, No. 3, pp. 766–799.

[13] Li Qi-na and Li Ting-hui, (2020), "Research on the application of Naive Bayes and Support Vector Machine algorithm on exercises Classification," Journal of Physics: Conference Series, Vol. 1437, No. 1.

[14] Anandarajan Murugan, Hill Chelsey & Nolan Thomas, (2019), "Classification Analysis: Machine Learning Applied To Text. In: Practical Text Analytics", Advances in Analytics and Data Science, Vol 2. Springer, Cham.

[15] Anantharaman Adita, Jadiya Arpit, Tulasi Chandana, Siri Saj , Nvs Adikar Bharath & Mohan Biju, (2019), "Performance Evaluation of Topic Modeling Algorithms for Text Classification," , IEEE, 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, pp. 704–708.
[16] Han Jiawei, and Kamber Micheline, (2006), "Data Mining: Concepts and Techniques", second edition, Amsterdam: Elsevier/Morgan Kaufmann.
[17] Danil Muhammad, Efendi Syahril & Sembiring Rahmat Widia , (2019),"The Analysis of Attribution Reduction of K-Nearest Neighbor (KNN) Algorithm by Using Chi-Square", Journal of Physics: Conference Series. Vol.1424 012004.
[18] Patra Anuradha & Singh Divakar, (2016), "A Survey Report On Text Classification With Different Term Weighing Methods And Comparison Between Classification Algorithms", International Journal of Computer Applications, Vol. **75**, No. 7.
[19] Akay Mehmet Fatih, (2017), "Support Vector Machines Combined With Feature Selection For Diabetes Diagnosis,", Elsevier, Journal of Electrical and Electronics Engineering, Vol. 17, No. 2, pp. 3219–3225.
[20] Karamizadeh Sasan, Abdullah Shahidan & Halimi Mehran, (2014), "Advantage And Drawback Of Support Vector Machine Functionality", IEEE, 2014 International Conference on Computer, Communication and Control Technology in Malaysia, pp. 63–65.
[21] Rustam Z., Nadhifa F. & Acar M, (2018), "Comparison of SVM and FSVM for predicting bank failures using chi-square feature selection", Journal of Physics: Conference Series. Vol.1108, 012115
[22] U˘guz Harun, (2011), "A Two-Stage Feature Selection Method For Text Categorization By Using Information Gain, Principal Component Analysis And Genetic Algorithm" Elsevier Knowledge-Based System, Vol. 24, No. 7, pp. 1024–1032.
[23] Wang You-wei & Feng Li-Zhou, (2018) "A New Feature Selection Method For Handling Redundant Information In Text Classification,", Frontiers of Information Technology & Electronic Engineering, Vol. 19, No. 2, pp. 221–234.
[24] Yin Chunyong & Xi Jinwen, (2017), "Maximum Entropy Model For Mobile Text Classification In Cloud Computing Using Improved Information Gain Algorithm,", Springer, Multimedia Tools and Applications, Vol. 76, No. 16, pp. 16875–16891.
[25] 20Newsgroups. http://qwone.com/ jason/20Newsgroups/.
[26] Reuters, http://konect.cc/networks/gottron-reuters/.
[27] Watan.pdf: https://sourceforge.net/projects/arabiccorpus/.
[28] Khalaf Zainab A. & Hassan Khadeija A., (2018), "Filtering Approach And System Combination For Arabic News Classification," Journal of Theoretical and Applied Information Technology, Vol. 96, No. 14, pp. 4491–4501.
[29] Unnikrishnan P., Govindan V. K., & Madhu Kumar, (2019), "Enhanced sparse representation classifier for text classification," Expert Systems with Applications, Vol. 129, pp. 260–272.
[30] Sabbah Thabit, Ayyash Mosab &Ashraf Mahmood, (2018), "Hybrid support vector machine based feature selection method for text classification," , International Arab Journal of Information Technology, Vol. 15, No. 3A Special Issue, pp. 599–609.

Encryption and Steganography a secret data using circle shapes in colored images

Zeena N. Al-kateeb[1]*, Muna Jaffer AL-Shamdeen[2] and Farah Saad Al-Mukhtar[3]

1,2 Computer Sciences Department, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq
3Computer Science Department, College of Science, Al-Nahrain University, Baghdad/Iraq
*Corresponding author's e-mail address: zeenaalkateeb@yahoo.com, zeenaalkateeb@uomosul.edu.iq

Abstract: Since the earliest times, people have used the encryption and hiding data to achieve a safe and reliable transfer of important data. This paper propesed a new method for encrypting important data based on the circular shapes information that extracted from the cover image, The encryption process is done using an update of a well known traditional method with simple calculations taking advantage of the coordinates of the center of the main circle as keys extracted from the cover image to reduce the number of keys exchanged between the sender and the recipient and to increase the level of security and confidentiality. The hiding process of the encrypted data is done in pixels that located in the circular areas of the cover image and in three forms of concealment, which providing second encryption, choosing the hiding form depending on the appearing sequence of the character in the text, which makes decode the secure data will be so difficult, the experiments showed that the proposed method was achieved excellent encryption and hiding depending on the coefficients Peak Signal to Noise Ratio analyses (PNSR), Mean Square Error (MSE), and other measurements. The proposed method achieves a complete data recovery ratio where it was Bit Error Rate BER=0.

1. Introduction

In recent years, the Internet has been considered an appropriate method for transmitting digital and multimedia data, where it can be considered a cheap and fast way to transfer multimedia, digital data and other files in different fields like the private sector, government, medical areas, and military[1].

The security weakness of data is the main disadvantage of using the internet, that is because any unauthorized users can be monitored the data, which prompted the use of steganography [2].

Steganography is a method of to protect communication and to reduce attack risking during transportation over communication media. Steganography was introduced with the example of "Prisoner's secret message" by Simmons in 1983. Generally can be used all types of files, like text, image, video, and audio as carriers for the steganography. However, a high redundancy medium is more suitable ,so the image and audio files are the ideal format that used for steganography [3].

The classic steganography is interested in including a secret message in the cover medium , to increase security level, the keys are used, that making remove or detect the embedded original data very difficult without knowing the key used in it.

Capability and security are important objectives of steganography, capacity is the amount of data that can be hidden in the cover media . Security is interested in making important data unknown either by an unauthorized computer or a person. The hiding information process in a steganography system starts with identifying a redundant bits in cover's media. By exchanging the data from hidden messages with these redundant bits, the creation of a stage done by the embedding process.

Steganography and cryptography are methods to protect or hide secret data. However, in respect, they are different than steganography hides data exists, but cryptography hides data meaning.

These techniques are important, that used for providing security of the network [4]. Many researchers suggested useing the chaotic system to increase the reliability of cryptography and steganography [5],[6],[7]. In this paper, a method used to combine steganography and cryptography in one system

2. Related work

In modern times, many researchers have proposed many different systems to encrypt and hide secure data. In [8] Juneja and Sandhu, suggest an amended steganography technique based on (least Significant bit) for images that impart the best security of information. It shows an embedding algorithm to hide encrypted messages in random and not neighboring locations of a pixel in the smooth region and edges of images. In the first step the secret message encrypted, then in the cover image the edges detected using amended detection filter. After that, the message's bits will be embedded in LSB of edges area pixels that selected randomly. This ensures that the eavesdroppers do not suspect that the media contains any hidden message and makes steganography techniques ineffective to properly guess the length of the secret message. Balvinder et al. shows a new LSB way for steganography technique, that enhances the existed least significant bit substitution to improve the security of hidden information [9]. An algorithm presented using an effective steganography technique to hide secret data in the images, that used an 8-bit random key for encrypting a secret message, also this key will be used for selected the pixel in the cover image, which encrypted data hide in it. In the first step, XOR operation applied between the secret message and 8-bit random key for encrypting the plain secret message, then, the encrypted message will be hidden in the least significant bit of selected pixels in the cover image. Finally, the 2nd least significant bit of each pixel and the 8-bit random key are used to choosing the pixel which stores the encrypted message by applying some operation. As compared to other techniques this method, hiding a large number of characters in the cover image of the secret message as well as it is more secure. Purnama and Rohayani produce ciphertext by modifying the Caesar cipher method, this is done by replacing the alphabet into two parts, the consonant alphabet was replaced with a consonantal alphabet and the vocals were replaced with the alphabet vocal too [10]. However, because of the frequency of the alphabet is seldom used in an Indonesian text, there are some alphabet consonants are not replaced. A ciphertext that can be read is obtained from the tested result, with this ciphertext, the message does not suspect by the cryptanalyst so, he does not attempt to solve the ciphertext. Singh, et al. suggest a novel method for data- hiding based on the least significant bit technique of digital images, this technique using a lossless data hiding technique [11]. To derive the stego-image, The LSB algorithm is performed in spatial domain in which the payload bits embedded into the LSB of the cover image. tavoli, et al. discus the multiple approaches of steganography in an image, the proposed algorithm capable of storing a large amount of information[12]. A desirable percentage of steganography was yield by combining the LSB approach with mixing the use of the application of a particular scan of an image with an appropriate mask and adding a step of encrypting with each, these steps decrease the odds of discovering the hidden data. Alhassan et al. combined the cryptography and steganography to supply a robust system that can be able to encrypt a secret message using RSA algorithm and the advanced least significant bit method is used to hide the message [13]. In the first step, the original message encrypted and then separated into P1 and P2 portions, the first portion (P1), XOR operation applied to it using the odd location and to (p2) using the even position of the LSB+1. Then to hide the XORed encrypted message, the Position of the LSB is used. AbdelWahab et al. present a comparison between two different techniques, the first one (LSB) with no compression and no encryption, and the second one, before applying LSB technique, the secret message is encrypted [14]. Furthermore, transform the image into a frequency domain using Discrete Cosine Transform, to develop the stego-image, the least significant bit is performed in the spatial domain where the payload bits are inserted into a cover image in the LSB, while the DCT algorithm is performed in the frequency domain in which transforming the stego-image to the frequency domain and the payload bits are inserted into the cover image.

In 2018, Krishnaveni and Periyasamy suggested a secure system that gives high security and changed high implanting ability image steganography utilizing Least Significant Bit insert alongside chaotic supply map, During this system lossless and invisible amendment within the image steganography [15]. Whereas, Ogras (2019) proposed a spatial domain steganography technique which used the Logistic map for generating chaotic bitstream and bitwise XOR operation which is utilizied to create a control bit [16].

3. The proposed algorithm:

The proposed algorithm is based mainly on the circular shapes available in the cover image as one of the geometric shapes to applying its steps fully and correctly, The cover image must contain at least one circular area. The algorithm finds the circular areas using circular hough transform as a first elementary step and assigns the centers of those circles if there is more than one circular area in the cover image we assign one of those circles as the main circle and used its center coordinates for the coding and decoding process. To reduce the keys which must be exchanged between the sender and the recipient, we have suggested that the middle-chain circle should be the main circle (MC), if the number of circles (NC), then MC= round (NC/2) for example, NC=3 then MC= round (3/2), MC=2.

The encoding process is done in two phases. The original text is encoded using the X-axis coordinate of the center of the main circle and X-axis coordinates of the pixel that will contain the hidden data, and then the resulting text is encoded using the Y-coordinate of the center of the main circle and Y-axis coordinate for the pixel that will contain the hidden data as a second phase. Then the encrypted text will be embedded in the cover image. Embedding of encrypted data is based on the Least Significant Bit (LSB) method, but in a manner that ensures a kind of distributed diffusion based on the pixel location. The byte of the secret encrypted message is hidden in one pixel only, embed three bits from the secret encrypted message in the red layer and three bits in the green layer and two bits in the blue layer. The method involves hiding the secret message bits in the layers of the cover image in three maps:

1. Blue Green Red (BGR): The first two bits of the letter are stored in the blue layer, followed by the three bits stored in the green layer and the remainder of the bits in the red layer
2. Green Red Blue (GRB): The first three bits of the letter are stored in the green layer, followed by the three bits stored in the red layer and the two bits remainder are stored in the blue layer.
3. Red Green Blue (RGB): The first three bits of the letter are stored in the red layer, followed by the three bits stored in the green layer and the two bits remainder are stored in the blue layer.

Determine the way which used to hide will be based on the storage location, if the remainder of the row multiplier in the column equals 0, the first way of embedding will be chosen, but if the remainder of the row multiplier in the column equals 1, the second way of embedding will be chosen, if the remainder equals 2 then the third way will be chosen for embedding.

The main objective of the work is to encrypted secret data, and then hides encrypt secret data in specific locations of the cover image, but in a different sequence each time which gives a type of hash that acts as another type of encryption.

The encryption process is designed to make the data incomprehensible by unauthorized persons, The encryption stage in our algorithm goes through two phases to get encrypted data, We relied on encryption in one of the easiest and most famous encryption systems, the Caesar method, it uses the substitution of a letter by another one based on the shift value (key), in our algorithm we use the X-axis coordinates of the center of the main circle is key1 and is Y-axis coordinates of the center of the main circle is key2.

3.1 Encryption The Secret Text Steps:

1.Enter the secret text, plain text (PT)

2.Choose the cover image (CovIm)

3.Find circles in color images (CA)

4.Determine the center coordinates of main circle CX, CY, to use those coordinates in the encryption process

5.Encrypt the secret text using Caesar method based on the center coordinates of the main circle such as the following:

•Cipher text1 (CT1) = (PT) + (CX)

•Cipher text 2 (CT2) = (CT1) - CY

6.Get encrypted text

7.Repeat step 5 until all plain text is encrypted

3.2 Embedding The Encrypted Secret Text Steps:

1.Enter the secret ciphertext (CT2)
2.Choose the cover image (CovIm)
3.Find circles in the cover color image (CA)
4. Store all coordinates of the points within the circular regions in N*2 array lets ACP, where N is a no. of points within the circular regions, ACP[i][1] contains the X-axis coordinates of i point within the circular region, ACP[i][2] contain the Y-axis coordinates of i point within the circular region
5.Convert a character of the secret ciphertext (CT2) to asci_code (AsciCT2)
6.Convert the asci_code for a character of the secret ciphertext (AsciCT2) to binary code (BCT2)
7.Select a map to hide the character, depending on the location which that contain the character based on the following:
If (ACP[i][1]* ACP[i][2])%3
- =0 the hiding map a character is Blue Green Red (BGR)
- =1 the hiding map a character is Green Red Blue (GRB)
- =2 the hiding map a character is Red Green Blue (RGB)

9.Embedding a cipher character from secret ciphertext in a determined pixel of a cover image
10. Repeat step 5_7 until all ciphertext characters are embedding
11. Get the stego image (StegoIm)

3.3 Extracting Embedding Secret Text steps:

1.Enter the stego image (StegoIm)
2.Find circles in stego color image (CA)
3. Store all coordinates of the points within the circular regions in N*2 array lets ACP, where N is a no. of points within the circular regions, ACP[i][1] contain the X-axis coordinates of i point within the circular region, ACP[i][2] contains the Y-axis coordinates of i point within the circular region
4.Extract the embedding cipher character depending on the location that contains the character based on the following:
If (ACP[i][1]* ACP[i][2])%3
- =0 the hiding map a character is Blue Green Red (BGR)
- =1 the hiding map a character is Green Red Blue (GRB)
- =2 the hiding map a character is Red Green Blue (RGB)

5. Repeat step 4 until getting all extract cipher secret text (ECT).

3.4 Decryption The cipher Secret Text Steps:

1.Enter the extract secret ciphertext (ECT).
2.Enter the stego image (StegoIm)
3.Find circles in stego color image (CA)
4.Determine the center coordinates of main circle CX, CY, to use those coordinates in the encryption process
5.Decrypt the secret text character using Caesar method based on the center coordinates of the main circle such as the following:
6.Decrypt text1 (DT1) = (ECT)- (CY)
7.Decrypt text (DT) = (DT1) +(CX)
8.Repeat step 5 until Decrypt all extract cipher secret text.
The following block diagram clear Encryption&Embedding process in the sender side and Extracting &decryption process on the receiver side of the proposed algorithm, Fig.1.

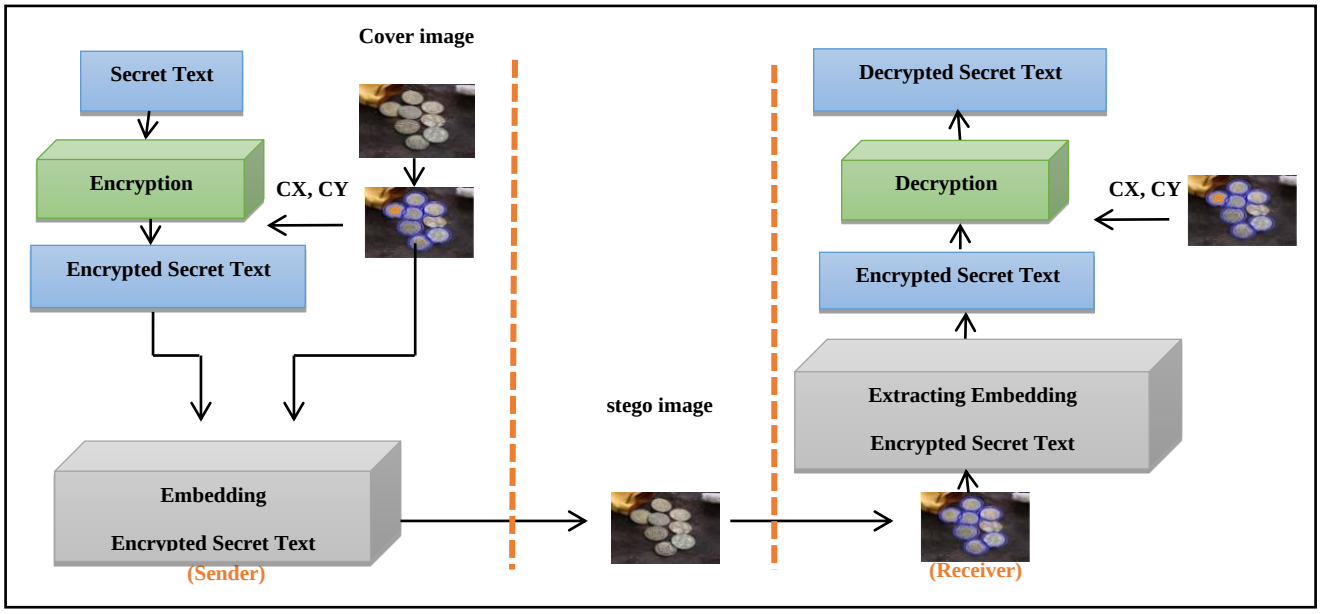


Figure 1. The block diagram of the proposed algorithm

4. Result and Analysis:

The proposed algorithm was built using Matlab 2019, and it has applied to a group of different size images to demonstrate the possibility of using it for all images size, we have chosen an images that contains one circular area and other images that containing more than one circular area to demonstrate the efficiency of defining circular areas in the proposed algorithm and the results were shown in table.1 which is below.

Table 1. The images before and after HidingMany metrics are used to demonstrate the efficiency of The Mean Square Error (MSE), the Peak Signal to Noise Ratio (PSNR) and the Signal to Noise Ratio (SNR), correlation coefficient (Corr), The Structural Similarity index (SSIM) are used to test the quality of image Steganography, Bit error rate (BER) are used to test retrieval data .

MSE represents the squared difference between the original image and the stego image, whereas PSNR represents a measure of the peak error [17],[18],[19].
The general form of PSNR, MSE, and SNR is as follows.

$$PSNR = 10 \log_{10}\left(\frac{C_{max}^2}{MSE}\right) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Where $C_{\_max}$ represents the maximum value in the image.

$$MSE = \frac{1}{MN} \sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij} - y_{ij})2 \dots\dots\dots\dots\dots\dots .. (2)$$

$$SNR = 10 \log_{10} \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij}^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}[x_{ij} - y_{ij}]\,2} \dots\dots\dots\dots\dots\dots\dots (3)$$

Where
M,N Represents the row and column of the image, $x_{\_(ij)}$ represents the cover image and $y_{\_ij}$ represents an image that contains hidden information.

The correlation coefficient is used for the purpose of comparing two images and noting the extent of their convergence. The best value for the correlation coefficient is to be close to one [20][21][22][23].

$$Corr = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij} - \bar{x})(y_{ij} - \bar{y})}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij} - \bar{x})2}\sqrt{\sum_{i=1}^{N}(\sum_{j=1}^{M}(y_{ij} - \bar{y})2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Where $\bar{x}$, $\bar{y}$ Represents the mean of image x and y and can be found by the following equation

$\bar{x} = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{x_{ij}}{NM}$ …….. (5) And $\bar{y} = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{y_{ij}}{NM}$…….. (6)

The Structural Similarity index (SSIM) is a method for calculating the similarity between two images and its points how much a stego image is similar to the original image[24] .
Bit error rate (BER) calculates the actual number of bit positions that are varied in the stego-image compared with the original image [25].

The results of the image metrics are shown in Table1 and Table2 give the values of MSE, Corr, SNR and PSNR, SSIM and BER of different types of images of different types of images.

Table 1 (Mean Square Error (MSE), correlation coefficient (Corr), Signal to Noise Ratio (SNR)) metrics for the proposed method

Image name	MSE	Corr	SNR
2f	0.00152582626809431	0.999999517163353	65.9138180628453
3f	0.000206063716456326	0.999999861813952	72.1752538005781
6f	0.00208308455731136	0.999999142735082	66.9183276780933
7f	0.00171821305841924	0.999999283277961	62.8273572062556
8f	0.00214650205761317	0.999999042575196	65.0787168613582
9f	0.00192655187074830	0.999999335151924	63.5526944962583

Table 2 (Peak Signal to Noise Ratio (PSNR), Structural Similarity index (SSIM), Bit Error Rate (BER) metrics for the proposed method)

Image name	PSNR	SSIM	BER
2f	71.6848146429595	0.999998290473524	0
3f	79.1728726584688	0.999999710432994	0
6f	71.7638992832398	0.999993563995055	0
7f	71.6458058943006	0.999988174265215	0
8f	70.5956350035664	0.999998567312157	0
9f	71.9830596275765	0.999960004761695	0

Examining the results in table 1, we note that the value of the MSE scale for all images is a small amount if it does not exceed the value 0.00214650205761317, which gives a clear impression on the quality of the algorithm, as we note that the value of Corr between the original image and the image after hiding is a value close to one Where was the lowest value 0.999999042575196, which indicates that the ratio of the image correlation before hiding and the image after hiding is a high correlation

For BER, its value equals to zeros for all images, which indicates that the hidden text is fully retrieved without any decrease. This is a clear indication of the quality and efficiency of the proposed algorithm.(If the value of BER is closer to zero it means the quality of the image is good).

Conclusion :

The idea of using circular area coordinates in the encryption and steganography processes in digital color images is an efficient and new idea at the same time, because of the availability of many images that include circular shapes such as football images, circular coin images, traffic images, and many others images that are not questionable or suspicious. The proposed algorithm encrypted secret text data using the simple and traditional Caesar method, but in an innovative way based on the coordinates of the central circular shape that the algorithm finds. This is followed by hiding or embedding encrypted data in specific locations within the color image using a method that provides a second encryption process depending on the location where the data will be embedded,this will be providing the possibility of encrypting the same character two different encryptions in the same message, which will be providing high security in the data transfer, The difficulty of decoding by unauthorized persons, as well as the full retrieval of that data at the receiving end of the communication. The experiments confirmed the efficiency and quality of the work.

Acknowledgments

The authors are very grateful to the University of Mosul/ College of Computer Science and Mathematics and Al-Nahrain University / College of Science for their provided facilities, which helped to improve the quality of this work.

References

[1]K. Joshi, S. Gill, and R. Yadav, 2018, "A New Method of Image Steganography Using 7th Bit of a Pixel as Indicator by Introducing the Successive Temporary Pixel in the Gray Scale Image", Journal of Computer Networks and Communications, (8),1-10, Article ID 9475142.

[2]P. Chandarana, P. Ahirao, 2018, "ADVANCED IMAGE STEGANOGRAPHY", International Journal Of Innovative Research in Information Security (IJIRIS), 5(7).

[3]A. S. Ansar, M. S. Mohammadi, and M. T. Parvez, 2019, "A Comparative Study of Recent Steganography Techniques for Multiple Image Formats", I.J. Computer Network and Information Security, 11(1) , 11-25.

[4]A. AL-Shaaby, T. AlKharobi, 2017,"Cryptography and Steganography: New Approach", TRANSACTIONS ON NETWORKS AND COMMUNICATIONS, 5(6), 25-38.

[5]S. F. Al-Azzawi and M. M. Aziz, "Chaos Synchronization of Nonlinear Dynamical Systems via a Novel Analytical Approach," Alexandria Engineering Journal, vol. 57, no. 4, pp. 3493-3500, Dec 2018.

[6]S. F. Al-Azzawi, et al., "Chaotic Lorenz System and it's Suppressed," Journal of Advanced Research in Dynamical and Control Systems, vol.12, no. 2, pp. 548-555, 2020.

[7]A. S. Al-Obeidi and S. F. AL-Azzawi, "Chaos Synchronization in a 6-D Hyperchaotic System with Self-Excited Attractor," TELKOMNIKA Telecommunication, Computing, Electronics and Control, vol. 18, no 3, pp. 1483-1490, June 2020.

[8] M. Juneja, P. S. Sandhu, 2013, "An Improved LSB Based Steganography Technique for RGB Color Images", International Journal of Computer and Communication Engineering, 2(4), 513-517

[9]B. Singh, S. Kataria, T. Kumar, and N. S. Shekhawat, 2014, "A Steganography Algorithm for Hiding Secret Message inside Image using Random Key", International Journal of Engineering Research & Technology (IJERT), 3(12).

[10] B. Purnama, H. Rohayani, 2015, "A New Modified Caesar Cipher Cryptography Method With Legible Ciphertext From A Message To Be Encrypted", Procedia Computer Science, 59, 195-204.

[11] A. K. Singh, J. Singh, H. V. Singh, 2015, "Steganography in Images Using LSB Technique", International Journal of Latest Trends in Engineering and Technology (IJLTET), 5(1), 426-430.

[12] R. Tavoli, M. Bakhshi, F. Salehian,2016, “A New Method for Text Hiding in the Image by Using LSB”, (IJACSA) International Journal of Advanced Computer Science and Applications, 7(4), 126-132.
[13] J. K. Alhassan, I. Ismaila, V. O. Waziri, and A. Abdulkadir, 2016, “ A Secure Method to Hide Confidential Data Using Cryptography and Steganography”, Federal University of Technology, Minna, Nigeria November, 28-30.
[14] O. F. AbdelWahab, A. I. Hussein, H. F. A. Hamed, H. M. Kelash, A. A. Khalaf, H. M. Ali,2019, “ Hiding data in images using steganography techniques with compression algorithms”, Telkomnika, 17(3), 1168-1175.
[15] N. Krishnaveni, S. Periyasamy, 2018, “Image steganography using LSB embedding with chaos ”, 118,505-509.
[16] H. Ogras, 2019, “An Efficient Steganography Technique for Images using Chaotic Bitstream” International Journal of Computer Network and Information Security, 11(2), 21.
[17] A.Sundar, V. Pahwa, C. Das, M. Deshmukh, N. Robinson, 2016, “A Comprehensive Assessment of the Performance of Modern Algorithms for Enhancement of Digital Volume Pulse Signals ", International Journal of Pharma Medicine and Biological Sciences, 5(1), 91.
[18] Y. Inan, 2018, “Assesment of the Image Distortion in Using Various Bit Lengths of Steganographic LSB ” , Near east University, Computer Engineering Department, Nicosia, TRNC, Mersin 10 Turkey, ITM Web of Conferences 22, 01026.
[19] E. Noroozi, S. B. M. Daud, A. Sabouhi, 2011, “Critical Evaluation on Steganography Metrics ”, In Advanced Materials Research,748, 927-931, Trans Tech Publications Ltd.
[20] E. A. Albahrani, 2017, "A New Audio Encryption Algorithm Based on Chaotic Block Cipher", Annual Conference on New Trends in Information & Communications Technology Applications - (NTICT 2017), 22-27, IEEE.
[21] Z. N. Al-Khateeb, M. F. Jader, 2020, “Encryption and Hiding Text Using DNA Coding and Hyperchaotic System,” Indonesian Journal of Electrical Engineering and Computer Science, 19(2)Aug 2020.
[22] Z. N. Al-kateeb , M. R. Al-Bazaz, 2019 ,“Steganography in Colored Images Based on Biometrics,” Tikrit Journal of Pure Science, 24(3), 111-117.
[23] Z. N. Al-Khatee , S J. Mohammed, 2020, “A Novel Approach for Audio File Encryption Using Hand Geometry,” Multimedia Tools and Applications,Mar 2020
[24] E. Margalikas, S. Ramanauskaitė, 2019, “Image steganography based on color palette transformation in color space ” , EURASIP Journal on Image and Video Processing 2019(1), 82.
[25] M. S. Hiwe, S. I. Nipanikar, 2014, “ An Analysis of Image Steganography Methods” , International Journal of Engineering Research & Technology (IJERT) , ISSN: 2278-0181, 3(2).

Review of Different Combinations of Facial Expression Recognition System

Abd_Almuhsen, F. Almudhafer[1], Zainab A. Khalaf[2]

1 Basrah University, College of Education for Pure Sciences, Department of Computer Science, Basrah, Iraq
2 Basrah University, College of Science, Department of Mathematics, Basrah, Iraq
1 abdalmudhafer@gmail.com 2 zainab_ali2004@yahoo.com

Abstract. The facial expression recognition (FER) system is a classifier system that attempts to recognize facial expressions based on the analysis of emotion behaviour on the face. The FER system can be implemented by using one classifier or combining multi feature extraction and/or multi classifiers. In general, FER is used with one classifier system to find the best label. Although a classification system is commonly used to find the most likely facial expression, it still produces substantial numbers of errors due to several factors that influence the FER result, such as data quantity, and environmental conditions (i.e. illumination and noise). Therefore, combined multi feature extraction methods and/or multi classifier systems are useful to avoid the single classifier errors. Multi feature extraction or a multi classifier system combination are used to take advantage of different system hypotheses to find an accurate result. This paper is a survey of the latest system combination techniques being used to enhance the classification performance in the FER system; the most recent studies are presented.

1. Introduction

The human face reflects internal feelings in an immediate time frame, due to its important role in human personal communication. After viewing the face, the identification of the person, sex, expression, etc. will be recognized. Thus, facial expressions provide sensitive signals about feelings, and play a major role in human interaction and nonverbal communication [1]. Although there is a wide range of possible facial expressions, psychologists have identified six fundamental ones (happiness, sadness, surprise, anger, fear and disgust) that are universally recognized. It is obvious that a system capable of performing automatic recognition of human emotions is a desirable task for a set of applications such as human-computer interaction, security, affective computing, etc. Figure 1 is an example of these six fundamental expressions in addition to a normal facial expression.

Figure 1: An example of basic facial expressions

Recently many papers have proposed the combination of more than one classifier when designing systems for pattern classification with high performance that work in a hybrid way. The reason caused the increased need in multi-classifier systems is the acknowledgement that the classic methods for implementing a pattern

recognition system that directed to select the best classifier, suffer from some basic drawbacks. The basic drawback is the hardness in selecting the suitable classifier for the classification task, unless a deep former knowledge on the data is available. In addition, using one classifier won't let the exploitation of the complete distinctive data to be encapsulated by other classifiers [1]. The vision behind a combined classifiers is to merge multiple machine learning technique, in order to improve the system performance as best as available. In other words, a combination approach basically consists by two, three or more classifiers. For example, the first classifier takes the inputted data and generates initial findings. The second one takes those findings as the input and produce the final results [2]. Results based aggregation can be used if more than two classifiers are used to have the best results that represent the collective opinion of the whole system [3]. In facial expression recognition systems, a combination is used with more than one approach that seeks the general goal of the highest accuracy that can be reached. Two main approaches to be discussed are the prior-combination and the post-combination [4], as shown in figure 2.

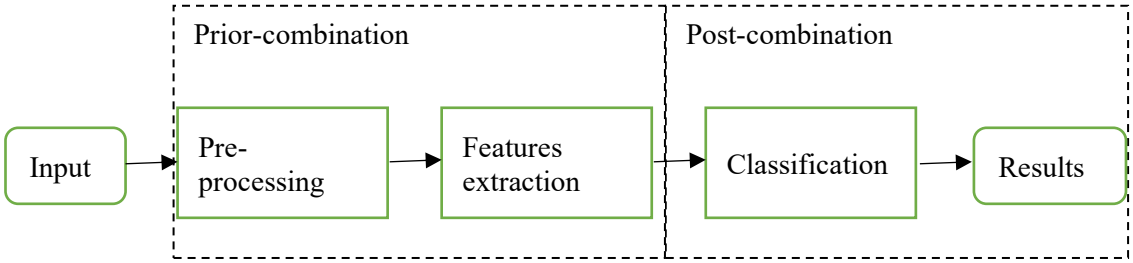


Figure 2: Typical system for FERS with combination

Prior-combination works on the feature extraction phase, which means that it might be a combined extraction method or a combined extracted features. Both ways will lead to a new set of features different to the initial sets. And many classifiers can be used to detect and extract features from the face such as local binary pattern, Hidden Markova model, the AdaBoost classifiers [5], Principle component analysis and eigenvectors [6]. The post-combination works on either enhancing the classifier results with the assistance of a second classifier or works on gathering more than one result in order to vote on the most frequent results to be chosen as the final result.

Figure 3: Typical Prior-combination System

Figure 3 demonstrates a typical scenario of the prior-combination that can occur when using two different extraction methods after the pre-processing step. This will lead to extraction of a two feature set that can be combined to provide a new set with more distinctive aspects.

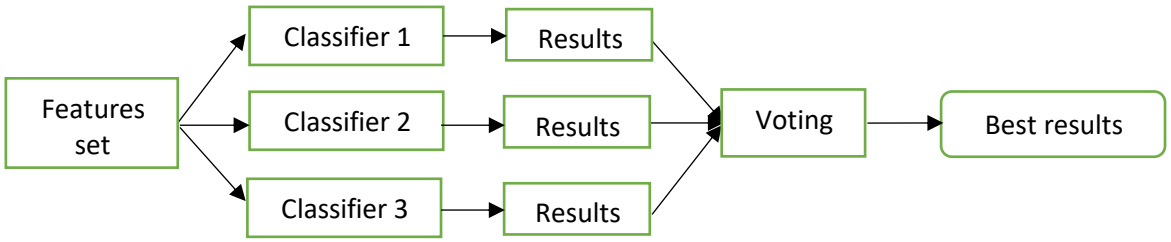


Figure 4: Typical Post-combination System

In a similar, but combined by using more than one classifier with the are used as a vote to select the best

result as detailed in figure 4. In another scenario, classifiers are combined to assess in handling part of the tested data, the first classifier will recognize part of the tested data according to a predefined condition and pass the remaining to the second classifier while the results are sent to the result aggregation pool, as detailed in figure 5. A study to compare classifier combination strategies was presented in Kuncheva's work [7]. Techniques like average, minimum, maximum, median, and majority votes were detailed in a theoretical way.

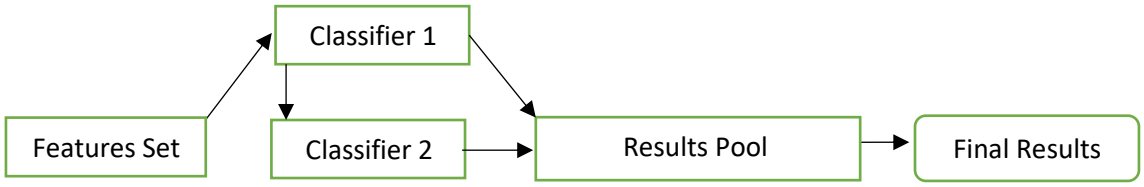


Figure 5: An example of Post-combination

The remaining parts of this paper are in four sections: Prior-Combination for features extraction and Post-Combination Classifier of recent studies are detailed in sections 2 and 3. Discussions and the conclusions follow in sections 4 and 5, respectively.

2. Prior-Combination for Features Extraction

Several studies, where more than one methodology was used, worked on extracting features with the best distinctive rate. For example, Sun et al. [8] proposed a robust facial expression recognition approach, directed to find the region of interest in the face to train a robust face-specific of Convolutional Neural Networks (CNN). By using the similar aspects between the facial areas within the ROI, an architecture stated aims to enhance the performance while predicting targets. The proposed architecture depends on deep learning to fine tune the new model of neural networks, the fine tune step is made according to a previously trained deep network to achieve the required performance. Also, the researchers worked to increase and improve the training process in the deep CNN by using data augmentation strategy. Kalsum et al. [9] designed a combination for a spatial bag of features (SBoFs) with spatial scale-invariant feature transform (SBoF-SSIFT). The SBoFs descriptor generated a feature vector with fixed length for all images used irrespective of their size and the SSIFT was designed by combining scale-invariant feature transform (SIFT) with speeded up robust transform features (SURF). The (SBoF-SSIFT) speeded up the transform process and enhanced the recognition ability of facial expressions. Because those features are independent of rotation, scale, translation, projective transforms, and when partial illumination might occur. For the recognition phase, K-Nearest Neighbor and Support vector machines were used. In another use of the bag of features combination, a feature extraction method was proposed by Sun and Lv [10] to be used in the facial expression recognition from a single image frame. The hybrid features used a combination of SIFT and deep learning features of different extraction level from the images, by using a trained CNN model. Also, Mahmood et al. [11] worked on facial variations and the complexity of appearance. In their study, the researchers attempted to improve the system accuracy by using Radon transform and Gabor wavelet transform. Facial detection was tested by the oval parameter method and facial tracking was achieved by implementing vertex mask generation. Radon transform and Gabor transform filters were applied to extract a variable set of features. Finally, self-organized maps with neural networks were used as the recognizing engine to measure the six facial expressions. Other researchers worked on using the distances between the facial landmarks, and on a triangular structure induced with three points in circumcenter, in center and centroid, which was considered as the geometric primitive to extract the required features. Information gained from those features were used for the task of discrimination of expressions using the Multilayer Perceptron (MLP) classifier in images containing facial expressions available in more widely known databases [12]. On the other hand, Sun worked on the idea of extracting

optical flow from the changes between the highest expression intensity in the face image and the normal expression face image as the temporal information of a facial expression. He used the grey image of the facial expression as the spatial information. Also, a multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) was presented to perform the spatiotemporal deep feature extraction with fusion from the static images [13].

3. Post-Combination Classifier System

On the other side of the combination diagram, the classifier is combined for the recognition task and to enhance the classifier performance to recognize features with high or low distinctively. Jain et al. [14] proposed combining sequential information while using Recurrent Neural Network (RNN) to spread gained information. The CNN model was used for the extraction of features in order to fix all of the CNN parameters and to be able to eliminate the regression layer. For the processing, after the image passes to the network, 200-dimensional vectors are extracted from the fully-connected layers. All vectors will go through a node of the RNN; and finally, all the nodes of the RNN returns a results of valence label. On the other hand, Vo and Lee [15] Presented a new methodology works on the hierarchical representation called hierarchical collaborative representation-based classification (HCRC). The researchers depend on a classifier designed with two stages, first is to use deep convolutional neural network (DCNN) which works on extracting distinctive features from the image. And the second stage is to combine the HCRC with a model of local ternary patterns (LTP) in order to enhance the classifier performance to be robust with noisy conditions. Kar and Babu worked on proposing a combination system with both features and recognition to classify facial expressions; the proposed system has three steps. In the beginning, ripple transform type II is used to find the features from the facial area in the image. This approach is efficiency and working with both edges and textures. The second step, a principal component analysis (PCA) along with linear discriminant analysis (LDA) method are used to get more discriminative features. In the final step, the recognition is done by the least squares variant of support vector machine (LS-SVM) with a radial based function (RBF) kernel as demonstrated in [16]. Another use of the CNN combination to figure out the video copy might happen by suggesting a 3D-CNN architecture with parallel condition to handle multi-class recognition by implementing one 3D-CNN in combination with multiple two-class classifier. Eventually, the 3D-CNN is proposed as the two classes' classifier in any of the two-class classification [17].

4. Discussions

In order to understand the differences between the methods used to design a FER system, the results of the methods mentioned are summarized as a compression in the table below in terms of method, type, database and accuracy. There are many databases used in the FER system, such as: the Japanese Female Facial Expression (JAFFE) [18], the Extended Cohen Kennedy (CK+) [19], the Facial Expression Database 2013 (FER-2013) [20], Media Understanding Group (MUG) [21] and many other databases can be found in the literature.

This study focuses on the effect of using combination system either prior- or post-combination system instead of using individual system in facial expression recognition. In order to show the advantages of combination system in enhancing the facial expression results, the experimental results of the recent related works are compared. For the CK+ data set, the researchers in [22] used SIFT shallow features to extract the facial features, and the accuracy was 79%. On the other hand, the study in [23] extracted the deep emotions features in feature extraction stage, the obtained accuracy was 80.07%. While, the study of [10] used the same data set by applying the combination of deep and shallow features. The results were 94.82% as accuracy. Furthermore, Zhang et al. [24] used SIFT features for feature extraction on the CK+ dataset, and their accuracy was 95.8%. Whereas, after combining a spatial bag of features (SBoFs) with spatial scale-invariant feature transform (SBoF-SSIFT) in research [9], the accuracy enhanced to 98.5% on the same dataset. For the mix of two datasets: MMI and JAFFE, Jain et al. [14] used CNN as classifier system and the obtained accuracy was 76.51%. Then, the researchers combined two deep learning classifiers CNN-RNN , and the accuracy enhanced to 91.20%. In addition, the researchers used the Hybrid CNN-RNN

model with the ReLU, and a significant performance achieved with 94.46%. Support Vector Machine (SVM) facial expression classification system has been used by Sohail and Bhattacharya [25]. In this study, fifteen different feature points used to face identification. The obtain accuracy of 92% on the JAFFE dataset, and 86.33% on the CK dataset. These facial expression classification results using SVMs as a classifier positively illustrate the strengths of SVMs for emotion recognition. In another study,
CNN with five convolutional layers to extract a feature vector was used by Ouellet [26], and then the vector feed to a SVM for classification. The researcher got 94.4% accuracy rate on CK+ Dataset. Ruiz-Garcia et al. [27] used for recognition a hybrid model combined CNN for feature extraction and SVM for classification. They tested it model on the CK+ dataset and achieved a classification performance of 95.87%. In recap, we can conclude that the combination of two (or more) types of systems or methods (either prior- or post-combination) could significantly enhance the overall result of facial expression detection. Consequently, the combination system can be considered as an indicator for promising results.

Ref No.	Year	Method Proposed	Combination Type	Databases	Accuracy
8	2020	ROI partitioning	Prior	JAFFE ,CK+ FER-2013, WCD	53.77% ,94.67% 40.13%,37.25%
9	2018	Hybrid feature descriptors	Prior	CK+, JAFFE	98.5%, 98.3%
10	2019	Combining deep and shallow features	Prior	CK+, JAFFE, MMI	94.13% , 48.9 %, 53.81%
11	2018	1D transform and GWT	Prior	CK , AT&T	86% , 83.7%
12	2020	Landmark triangulation	Prior	CK+,JAFFE, MMI, MUG	99.08%, 97.18% 96.87%, 97.26
13	2019	Deep spatial-temporal feature fusion	Prior	CK+, MMI, RaFD	98.38% , 99.17% , 99.59%
14	2018	Hybrid deep neural networks	Post	JAFFE, MMI	94.91% , 92.07%
15	2018	Hierarchical collaborative representation	Post	AR , Yale	99.9 % , 99.5%
16	2019	Ripplet II + LS SVM	Post	JAFFE , CK+	99.46 , 98.97%
Ref No.	Year	Method Proposed	Combination Type	Databases	Accuracy
8	2020	ROI partitioning	Prior	JAFFE ,CK+ FER-2013, WCD	53.77% ,94.67% 40.13%,37.25%
9	2018	Hybrid feature descriptors	Prior	CK+, JAFFE	98.5%, 98.3%
10	2019	Combining deep and shallow features	Prior	CK+, JAFFE, MMI	94.13% , 48.9 %, 53.81%
11	2018	1D transform and GWT	Prior	CK , AT&T	86% , 83.7%
12	2020	Landmark triangulation	Prior	CK+,JAFFE, MMI, MUG	99.08%, 97.18% 96.87%, 97.26
13	2019	Deep spatial-temporal feature fusion	Prior	CK+, MMI, RaFD	98.38% , 99.17% , 99.59%
14	2018	Hybrid deep neural networks	Post	JAFFE, MMI	94.91% , 92.07%
15	2018	Hierarchical collaborative representation	Post	AR , Yale	99.9 % , 99.5%
16	2019	Ripplet II + LS SVM	Post	JAFFE , CK+	99.46 , 98.97%

5. Conclusions

This paper is prepared to review the latest facial expression recognition system studies which worked on using a combination of either features extraction or the recognition of expressions. FERS is used in many

applications such as entertainment, medical, security, mental disorders and many other fields. Therefore, it is important to always have a high accuracy in recognizing facial expression. Multiple studies have been briefly reviewed in the previous sections to build a comprehensive view of recent theories proposed to enhance system performance. Prior- and Post-combination techniques were mentioned, either by using more than one method to combine results or features, or by enhancing the classifiers effectiveness and performance or the features extractor by the assistance of another one. In recap, different algorithm produce different results due to the robustness of the algorithms that are vary from one algorithm to another. Hence, using combination system gives promising results comparing with the individual system due to the combination system takes the advantages of different approaches to produce reliable result.

References

[1] F. Roli, G. Giacinto, and G. Vernazza, (2001, July). Methods for designing multiple classifier systems. In International Workshop on Multiple Classifier Systems (pp. 78–87). Springer, Berlin, Heidelberg.

[2] M. Mohandes, M. Deriche and S. O. Aliyu, "Classifiers Combination Techniques: A Comprehensive Review," in IEEE Access, vol. 6, pp. 19626-19639, 2018.

[3] A. M. Chacko, & Kumar, K. A. (2016). Offline Malayalam Character Recognition: A Comparative Study Using Multiple Classifier Combination Techniques. In Information Systems Design and Intelligent Applications (pp. 69-77). Springer, New Delhi.

[4] Z. A. Khalaf, (2015). Broadcast News Segmentation Using Automatic Speech Recognition System Combination with Rescoring and Noun Unification, PHD thesis, Universiti Sains Malaysia, Malaysia.

[5] W. M. Sanjaya, , D. Anggraeni, A. Juwardi, & M. Munawwaroh, (2018, September). Design of Real Time Facial Tracking and Expression Recognition for Human-Robot Interaction. In Journal of Physics: Conference Series (Vol. 1090, No. 1, pp. 012044). IOP Publishing.

[6] M. Azriansyah, N. Hartuti, M. Fachrurrozi, & B. A. Tama, (2019, March). A Study about Principle Component Analysis and Eigenface for Facial Extraction. In Journal of Physics: Conference Series (Vol. 1196, No. 1, pp. 012010). IOP Publishing.

[7] L. I. Kuncheva, (2002). A theoretical study on six classifier fusion strategies. IEEE Transactions on pattern analysis and machine intelligence, 24(2), pp. 281–286.

[8] X. Sun, P. Xia, L. Zhang, and L. Shao, (2020). A ROI-guided deep architecture for robust facial expressions recognition. Information Sciences. v 522, pp (35-48). Elsevier

[9] T. Kalsum, S. M. Anwar, M. Majid, B. Khan, and S. M. Ali, (2018). Emotion recognition from facial expressions using hybrid feature descriptors. IET Image Processing, 12(6), 1004–1012.

[10] X. Sun, and M. Lv, (2019). Facial expression recognition based on a hybrid model combining deep and shallow features. Cognitive Computation, 11(4), 587–597.

[11] M. Mahmood, A. Jalal and H. A. Evans, "Facial Expression Recognition in Image Sequences Using 1D Transform and Gabor Wavelet Transform," 2018 International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, 2018, pp. 1-6.

[12] A. Nandi, P. Dutta, and M. Nasir, (2020). Recognizing human emotions from facial images by Landmark Triangulation: A combined circumcenter-incenter-centroid trio feature-based method. In Algorithms in Machine Learning Paradigms (pp. 147–164). Springer, Singapore.

[13] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, (2019). Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recognition Letters, 119, pp. 49–61.

[14] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, (2018). Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters, 115, pp. 101–106.

[15] D. M. Vo, and S. W. Lee, (2018). Robust face recognition via hierarchical collaborative representation. Information Sciences, 432, pp. 332–346.

[16] N. B. Kar, K. S. Babu, A. K. Sangaiah, and S. Bakshi, (2019). Face expression recognition system based on ripplet transform type II and least square SVM. Multimedia Tools and Applications, 78(4), pp. 4789–4812.

[17] J. Li, H. Zhang, W. Wan, and J. Sun, (2018). Two-class 3D-CNN classifiers combination for video

copy detection. Multimedia Tools and Applications, pp. 1–13.
[18] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, “The japanese female facial expression (JAFFE) database,” in Proceedings of third international conference on automatic face and gesture recognition,1998, pp. 14–16.
[19] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews ,“The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010, pp. 94–101.
[20] P. L. Carrier, A. Courville, I. J. Goodfellow, M. Mirza, & Y. Bengio (2013). FER-2013 face database. Universit de Montral.
[21] N. Aifanti, C. Papachristou, & A. Delopoulos (2010, April). The MUG facial expression database. In 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10 (pp. 1-4). IEEE.
[22] R. Azhar, D. Tuwohingide, D. Kamudi & N. Suciati (2015). Batik image classification using sift feature extraction, bag of features and support vector machine. Procedia Computer Science, 72, 24-30.
[23] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera and D. Manocha, "Learning Perceived Emotion Using Affective and Deep Features for Mental Health Applications," *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Beijing, China, 2019, pp. 395-399.
[24] T. Zhang, W. Zheng, Z. Cui, Y. Zong and Y. Li, Spatial-Temporal Recurrent Neural Network for Emotion Recognition, IEEE Transactions on Cybernetics, arXiv:1705.04515, Issue: 99, pp. 1-9, 2018
[25] A. S. M. Sohail. & P. Bhattacharya (2011). Classifying facial expressions using level set method based lip contour detection and multi-class support vector machines. International Journal of Pattern Recognition and Artificial Intelligence, 25(06), 835-862.
[26] S. Ouellet (2014). Real-time emotion recognition for gaming using deep convolutional network features. arXiv preprint arXiv:1408.3750.
[27] A. Ruiz-Garcia, Elshaw, M., Altahhan, A., & Palade, V. (2018). A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. *Neural Computing and Applications*, *29*(7), 359-373.

Convert Arabic Letters Voice into Gesture

Shaker K. Ali1 Sabreen k. Saud2

1Computer sciences and mathematics college, University of Thi_Qar, Thi_Qar, Iraq.
2College of Eduction for Pure, University of Thi_Qar, Thi_Qar, Iraq.
1 shaker@utq.edu.iq and 2 sabreenkhalid55@gmail.com

Abstract: This paper suggest approach to solve the problem of social communication between blind and dumb by converting voices of 28 Arabic letters (أ,…….,ي) into gesture (images) by extraction features by using Mel-frequency Cepstral coefficients (MFCC)and classify the types of letters by using; J48, KNN, and Naive byes (NB). Several features are extracted from speech voice of Arabic letters voices. The dataset collected by recorded voices from twenty different persons, each person recorded ten voices for each twenty eight letters so the total dataset are 5600 voices (200 voices for each 28 letters). Mel-frequency Cepstral coefficients are extracted from 5600 voices of letters which convert the voices into a signal and extract features vector to classify later by using J48, KNN and NB algorithms, which may vary in time or speed signals. The experimental results shows that the best accuracy of speech recognition algorithm by using the J48 algorithm with a performance ratio of 100% while KNN is the 94.023% and Naive byes is the 20.012%.
Keywords: Mel frequency cepstral coefficients (MFCC), Vector quantization, J48, KNN and NB ,K-mean.

1.Introduction

Communication is the way of exchanging thoughts, opinions, information, or messages among the people by writing, speech, or signs. Communication bridges the gap between the people. There are different ways to communicate, where communication is usually oral among people i.e. talking to each other while dumb people cannot communicate with others as ordinary people do. They cannot speak, while the people who are deaf are able to speak, but unable to hear and the blind are unable to see but they can speak and hear the voice [1]. From the perspective of biology, sound can be considered as a signal that contain many or single tone which emanate from living organisms which own sounding member utilized to communicate between the individuals or other genus which express what is wanted by an action or speech and referred to as sense which result from such vibrations hearing. Voice is considered as the base for various experiences which are obtained via the individuals and the sound speed in the normal antenna's centre is determined at 343 meters for each second or 1224 kilo-meters for each hour. The sound's speed is associated to material's density and the hardness factor related to materials where sound moves [2]. Voice can be considered as signal related to infinite information. Speech signal's digital processing is of high importance for high-speed in addition to the adequate automatic voice recognition technologies. Today, it is utilized for disabled individuals, telephony military, and healthcare [3] There are 2 areas related to the voice recognition: speaker recognition and speech recognition, however the research was limited to focus on the field of speech recognition [4].

*1.1 Speech recognition***:**

The main way of communication between individuals is speech, also it is considered as the most effective and natural approach to exchange information between humans [5]. Speech Recognition, also referred to as (computer speech recognition or automatic speech recognition) can be defined as the capability of programs or machines in identifying phrases and words from spoken languages and convert them into certain format that is readable by machines. The major aim of this method is to develop the systems and methods for inputting the speech to the machines [6]. Speech is considered as distinctive signals that carry and convey many levels related to knowledge sources, non-linguistic as well as linguistic information.

Human speech is defined as complicated acoustic wave which is the result of the output coming from the effort of a speaker. Speech consist of sentences which contain words [7].

1.2 Types of Speech recognition

Speech can be defined as vocalization (speaking) related to words or single word which are representing single meaning. Utterances could be multiple sentences, single sentence, many words, or one word [10].

- Isolated Word: Typically, the isolated word recognizer demands each one of the utterances to have quiet on the two sides related to sample window. It does not indicate that it does accept single words, yet it does demand one utterance at the same time. This can be applied for conditions in which the users are demanded for providing just single word commands or responses, yet it is extremely inefficient for multiple word inputs. Also, it can be implemented easily due to the fact that the boundaries of words are understandable and words have tendency towards being plainly pronounced and that is the main benefit of such type [10]
- Connected Word also means one word received and analyzed by the system at the same time in the same way as isolated words but here the silence between the words is reduced so that they appear to be connected or as intermittent sentence.[11]
- Continuous speech the sentences used in this type of speech are considered to be the most difficult to apply in speech recognition. It is difficult to define the limits of each word on the one hand and the lack of precision in the pronunciation of words when they are in a sentence. [11]
- Spontaneous speech This speech type is no rehearsed and it is considered to be natural. ASR system with spontaneous speech must have the ability of handling various natural speech features like the words which are run together in addition to the minor stutters. Spontaneous (un-rehearsed) speech might consist of nonwords, mis-pronunciations, and false-starts [10].

1.3 Human Speech Production

Speech is a tool of communication. The speech production is considered as day-to-day mean in humans, also it is considered to be an uncomplicated mechanism, yet there is a high complexity in its internal approach. The first stage of speech production is breathing, since it includes 2 processes, the first one is to inhale, while the other is to exhale. Throughout the inhaling process, the air will enter the lungs, while in exhaling process, the air is going to flow out of organism. Throughout the exhaling process, the air is going to flow out through lungs, trachea, larynx, vocal folds, mouth, lips, nasal cavity, and so on. Articulatory movement is the movement of such organs. The articulatory movement control can be referred to as the motor control, also it is done via the brain [4]. The vocal cords in their normal position are open and approaching each other in case of talking according to the sound to be issued [5].

1.4 Research problem

According to statistics provided by the World Health Organization, there are 285 million people around the world suffer from blindness, one million are dumbed and many suffer from other physical disabilities.[1] The problem is how to help people with physical disabilities communicate easily with one another (blind and dumb). This research focuses on this problem and tries to develop a new system that enables the blind and dumb communicate with each other by turning speech into a signal.

1.5 Literature Review

Parwinder Pal Singh, Pushpa Rani 2014 [10] this study has introduced a method for extracting the features for the speech signals related to spoken words with the use of MFCC. It is considered as non-parametric frequency domain method that operates on the basis of the human auditory perception system. Initially, all the voice samples related to the words will be taken as input and through the use of praat tool denoise all the samples. After that, the coefficients will be extracted via the use of MFCC since such coefficients are collectively representing short term power spectrum that is related to the sound. This study effectively denoised input samples in the same time as extracting MFCC coefficients, also it took int account the Delta energy function and concluded that there is a possibility of increasing the MFCC coefficient based on the

requirements. Acceleration and velocity could be added for extracting twelve MFCC coefficients. The features have been extracted on the basis of the information which has been involved in speech signal. A file of extensions (.wav) has been used to store the extracted features.

Anjali Bala, Abhijeet Kumar and et al 2010 [32] Voice is considered as a signal related to infinite information. The digital processing regarding the speech signals is of high importance for adequate and high-speed recognition technologies. Today, it is applied by disabled individuals, telephony military, and healthcare, thus, digital signal processes like feature matching as well as feature extraction are the most recent issues for studying the voice signals. For the purpose of extracting significant information from speech signals, making decisions related to processes, in addition to obtaining results, data should be manipulated and analyzed. The major approaches which have been applied to extract features related to signals is finding MFCCs, since they are considered to be collectively representing short-term power spectrum related to the sound, on the basis of the linear cosine transform that is related to log power spectrum on non-linear mel scale of frequency. This study present MFCC for extracting features as well as DTW for comparing test patterns

Ms. Rupali S Chavan, Dr. Ganesh. S Sable 2013 [14] Speech can be considered as a major and a natural communication form in humans. There are a lot of factors associated to speech, such as speaker identification, speech synthesis, speech recognition, speech verification, and so on. The main aim of this research is studying the system of speech recognition with the use of HMM. The main aim of speech recognition is determining the present speech on the basis of spoken information. HMM is used for pattern training, MFCC is used for feature extraction. It is indicated that MFCC is applied majorly for speech's feature extraction since it is robust to noise, while HMM is optimum in modeling methods since it increases the speed and preciseness of recognition.

Om Prakash Prabhakar and Navneet Kumar Sahu 2013 [31] Speech is defined as the major communication mode in humans. Nowadays, the technologies of technologies are provided commercially for un-limited, yet wide range of tasks. Such technologies are enabling the machines to reliably and correctly respond to the voices of humans, and offer significant services. This study provides a summary related to the main technological perspectives related to the major development of speech recognition as well as providing overview method created in each one of the speech recognition stages. The basics related to speech recognition have been discussed, also its new developments have been examined. The AST system's performance depends on adopted feature extraction approach and speech recognition method for specific language is put to comparison. Recently, the demand for speech recognition studies has increased. The effectiveness of HMM method accompanied by MFCC features is more appropriate for such demands and provide optimum recognition results. The approaches are going to allow creating more powerful systems, used worldwide in the future.

K. M. Shiva Prasad1*, G. N. Kodanda Ramaiah2 ,et al 2017 [13] . human speech is defined as distinctive signal that carry and convey multiple levels related to the knowledge source, non-linguistic and linguistic information. Speech signals might be considered as information bearing signals that are evolving as functions related to single independent variable such as time. Speech can be defined as complicated acoustic wave that result from the output related to the effort of speaker. Speech consists of sentences which are made of words. Generally, words are composed of phoneme sequences referred to as syllables. The speech analysis is of high importance in speech recognition and synthesis. Speech analysis is also referred to as the feature extraction. There are various approaches of speech analysis, each one of them have certain advantages and disadvantages, no single approach is defined as optimum for speech analysis or recognition. Speech analysis or speech signal front ends enable the extraction of speech features. LPC, PLP, and MFCC are majorly applied techniques of feature extraction in the speech analysis. The speech's nonlinear nature made the selection of LPC not excellent for speech estimation, MFCC and PLP are derived from logarithmically spaced filter banks combined with human auditory system, thus they have more optimum response in comparison to LPC. Rasta PLP approach is uncomplicated computationally effective and have high robustness for studying state spectral factors.x cv

Dr. Jaffar Alkhier 2017 [8] The speech recognition is one of the most modern technologies, which entered force in various fields of life, whether medical or security or industrial techniques. three systems have been created for speech recognition. They differ from each other in the used methods during the stage of features extraction .While the first system used MFCC algorithm, the second system used LPCC algorithm, and the third system used PLP algorithm. All these three systems used HMM as classifier.

2. Proposed System

The proposed system used Cross-Validation as statistical method, which performs in two parts; training and testing part at the same time. Where the testing part used 10% of the recorded voices and 90% for training. The training part used 5040 voices and the testing part used 560 voices. The system consist from four stages, the first stage is the recording of sounds (dataset collection) and the second stage is feature extraction using Mel Frequency Cepstrum Coefficient algorithm which contains six steps; (the first step Pre-emphasis , the second step Framing , the third step Hamming windowing, the fourth step Fast Fourier Transform, the fifth step Mel scale filter bank and the sixth step Discreet Cosine Transform). The third stage is vector quantization with the use of K-means algorithm. The fourth stage is the classification stage using the following algorithms (J48, KNN and Naive Bayes). The Figure.1 shows the flowchart of proposed system.

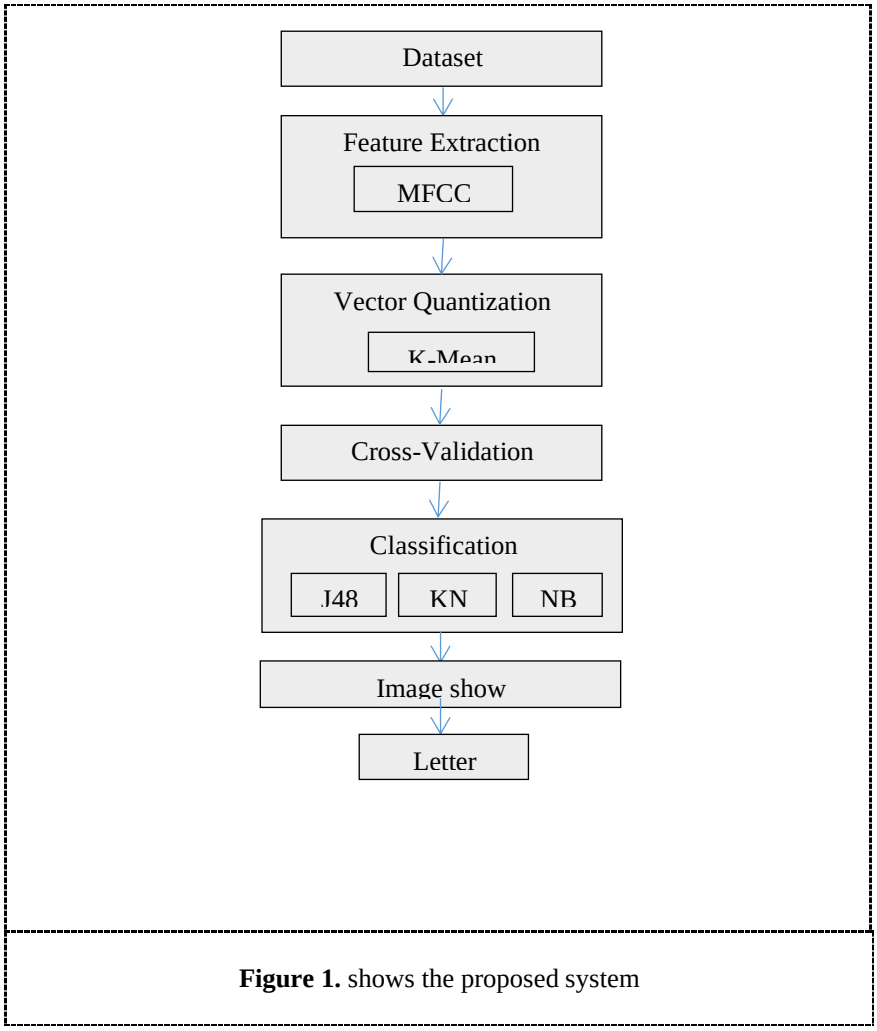


Figure 1. shows the proposed system

2.1. Dataset

Dataset collected from 20 different persons (males and females) by recording voices from 20 different persons, each person recorded 10 voices for each 28 letters so the total dataset are 5600 voices (200 voices for each 28 letters), which have been recorded in same environment. All voices signals have been acquired under various conditions, like the record time length, and sound amplitude level. The recorded voices are stored in format ".WAV" extension. Postulates to user for choosing any sample of the speech for testing from recorded dataset.

2.2. Feature extraction

The feature extraction is the basic portion of the system of speech recognition the feature extraction is the core of the system. Its function is to extract the features from input speech (signal). Extraction features compress the amount of input signal (i.e. the vector) with no causing of damages to the speech signal power [12]. At this stage, the speech signal is converted to a sequence of characteristic (feature vector) that represent information which is stored in the spoken speech. An important characteristic of feature extraction phase is the suppression of information which does not matter in order to properly classify such as information about the speaker and information about the transmission channel such as the telephone. Extraction features play an important role in speech recognition. It is very difficult to get data from speech signal [8].

There are several techniques that are used to do the job such as MFCC (Mel Frequency Cepstral Coefficients), RASTA filtering, LPC (Linear Predictive Coding), and PLDA (Probabilistic Linear Discriminate Analysis [13].

2.2.1 MFCC Feature Extraction

This technique is a very common technique in extracting the distinctive features of sound in speech recognition systems due to the accuracy of its results and the ability to partially eliminate the noise of the signal [14] As well as the speed of application and less complex and more effective under different circumstances [15] In this technique, the human hearing process is approached, ie, an attempt to extract the signal characteristics in a manner consistent with the human hearing mechanism, since the human ear is sensitive to frequencies that are less than 1,000 HZ and weak to frequencies higher than 1,000 HZ [11] . The reason for the design of such filters is that the human ear is not sensitive to high frequencies and therefore can reduce the number of filters characteristic of these frequencies [16]. It has been discovered that MFCC is commonly utilized for extracting speech features due to its robustness to noise [17].

MFCC is used due to the following reasons [13].

- MFCC is the most significant features that are needed between various speech application types.
- Provides highly accurate results for cleaning speech
- MFCC may be considered as standard features in speech recognition systems

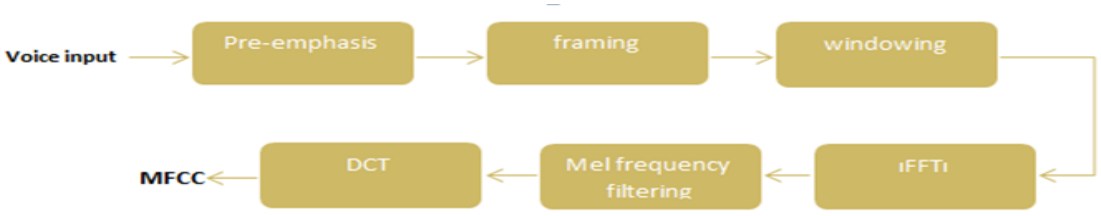


Figure 2. Mel-Frequency Cepstrum Co-efficients

In the MFCC, the voice signal passes through the following stages:

Step 1: **Pre-emphasis**

In the pre-emphasis, undesirable frequencies are produced from the environment during the sound recording process. These frequencies are low frequencies that are not considered to be spoken by the person and should therefore be omitted from the signal [4]. Each value in the speech signal is reassessed using the following Equation (1), [16].

$$y(n) = x(n) - a * x(n-1) \quad (1)$$

where $y(n)$ is the output signal, the value of a is usually between [0.9 -1.0] and $x(n-1)$ is the input signal.

Step 2: Framing

The speech signal is a constantly changing signal, so a framing process must be applied. The framing process is divide the signal into several sections each section called frame so that every one of the frames may be analyzed independently and in a short time rather than analyzing the whole signal. The signal cannot be handled once as this may cause unsatisfactory results [13]. The length of every one of the frames ranges between 20 and 40 with an overlap equal to half or one third of the size of the frame for easy transition from one frame to another [18].

Step 3: **Hamming windowing**

At this stage each of the above frames is multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame of the signal and eliminate interruptions at the edges and distortion in the signal [7]. The Hamming window is represented as in Equation 2. If the window is defined as $W(n), 0 \leq n \leq N-1$. where ; N = number of samples in each frame, Y[n] = Output signal, X (n) = input signal, W (n) = Hamming window, then the result of windowing signal as in Equation (2) [3].

$$W(n) = 0.54 - 0.46 * cos(2\pi n/(N-1)) \quad (2).$$

$$Y(n) = X(n) \times W(n) \quad (3)$$

Step 4**: Fast Fourier Transform (FFT)**

To convert each frame of N samples from time domain into frequency domain [3]. FFT is a fast approach of Discrete Fourier Transformation (DFT), on a specific set of N samples, the FFT Equation(4) is given as Where k= 0, 1, 2 ….. N-1

$$\mathrm{F(u)} = \sum_{k=0}^{N} f(k)\ e^{\frac{-j2\pi uk}{\mathrm{N}}} \quad (4)$$

Step 5: **Mel scale filter bank Frequencies**

It is a set of trigonometric signals, to calculate the filter banks, triangular filters are used and the frequency in Hertz (f) each frequency filter is given by Equation(5) [7].

$$Mel(f) = 2595 * log10\left(1 + \frac{f}{700}\right) \quad (5)$$

These filters exhibit linear behavior for low frequencies and logarithmic behavior for high frequencies. The reason for the design of such filters is that the human ear is not sensitive to high frequencies [11].

Step 6: Discrete Cosine Transform (DCT)

DCT is the procedure of converting Mel scale spectra into time-based domains. This process's result is referred to as the MFCC. The group of the obtained coefficients is referred to as the acoustic vector. Which means that until this step, audio inputs are converted to a stream of the acoustic vectors that will subsequently produce the group of inputs for algorithms of classification [17]. Given by the Equation(6) [19].

$$\mathrm{Cn} = \sum_{k-1}^{k} (logDk)\ \cos\left[\mathrm{m}(\mathrm{k} - \frac{1}{2})\frac{\pi}{k}\right] \quad (6)$$

where m = 0, 1…k-1,Cn represents the MFCC and m is the number of the coefficients here m=13 so, total number of coefficients extracted from each frame is 13.

2. 3. Vector quantization

Vector quantization can be defined as a conventional method of quantization from signal processing and the data compress in spatial domain. Due to the fact that it's one of the lossy techniques, therefore maintaining the quality of the image and the ratio of compression is a complicated task. Which is why, the code-book storing image data has to be designed in the best way [20]. K-means is utilized for the optimization of code-book [21].

2.4. Classification:

Classification is a method to extract the data and the division of this data into classes or specific groups in advance. It is a way to learn under the supervision of training requires disaggregated data. Classification is utilized for classifying every item in a dataset to one of the specified set of classes [22].There are many classification algorithms in this thesis. Such as J48, KNN and Naive Bayes algorithm.

2.4.1. J-48 algorithm

J-48 is a simple of C4.5 classification decision tree. It produces a binary tree. The method of the decision tree is of highest usefulness in the tasks of classification. With this approach, a tree is created for the sake of modelling the procedure of the classification. As soon as the tree has been constructed, it's implemented on every dataset tuple and produces the result of classification for the tuple. Throughout tree construction, J-48 overlooks missing values, which means that that item's value may be projected according to the information available on attribute values for the rest of the records. The main concept is dividing data to range according to that item's attribute values, which can be seen in training sample. J-48 can classify through decision trees or through the rules that are produced from those trees [22]. The J48 Algorithm as following:

- Basic algorithm, the tree is produced in a recursive top-down divide-and conquer way. First, every one of the training samples is at root. Attributes are categorical (if continuous-valued, they're previously discredited). The samples are recursively partitioned according to the chosen attributes. The test attributes are chosen based on statistical or heuristic measures (like the information gain)
- Conditions to stop partitioning: every sample for a certain node is part of the same class. There aren't any examples left [13].

4.2.2. K-Nearest Neighbour classifier (K-NN)

K-NN is a very significant non-parameter algorithm in the area of pattern identification, and it is one of the supervised learning predictable classification algorithms [23]. Amongst different approaches of the supervised statistical pattern identification, the rule of the Nearest Neighbour has achieved constantly higher performance, with no prior assumption on distributions from where training samples are derived. A new sample is categorized through the calculation of distance to nearest training sample [24]. KNN classifier expands this concept through taking k nearest points and assigning the majority class. For the sake of simplifying the issue it's usually fixed to odd a number (usually 1, 3 or 5) in order to leave no chance for ties. Bigger values of k are helpful in the reduction of the noisy points' effects in the training dataset, and k selection is usually carried out via cross-validation [25].The following steps give K-NN algorithm:

- Let k be a positive integer
- Compute distance d(x, xi) for each i=1,2,3….n where 'd' is Euclidean distance.
- Sort the samples based on the computed distance values.
- Select the heuristically optimum KNN based according the value of the RMSE which is performed with the method of the cross validation.
- Compute the inverse distance weighted mean with the k-nearest multi-variate neighbours.

4.2.3. Naïve Bayes classifier (NB)

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities. The algorithm tends to perform well and learn rapidly in various supervised classification problems, this property makes it suitable for datasets that are large [22]. The NB main idea is based on the so-called Bayesian theorem which is more suitable for inputs with high dimensionality. The model is called naive because it assumes that the attributes are conditionally independent of each other given the class. This assumption gives it the ability to compute Bayesian formula probabilities from a rather small training dataset [9]. Bayes classification is related to obtaining two probability types- Prior and posterior. The prior probability is not dependent on any data and is related only to the identification that a tuple is part of a specific class regardless any other data. Posterior probabilities are obtaining the probability of the tuple that belongs to a particular class according to some data. It's a conditional probability[26]. Bayesian theorem can be represented as:

$$P(H|T) = \frac{P(T|H)P(H)}{P(T)} \qquad (7)$$

"T" represents a tuple from a data-set D and H is the likelihood of a certain tuple T falling under a certain class. $P(T|H)$ and $P(H|T)$ represent the posterior probabilities denoting the conditional evaluations of T on H and H on T, respectively. $P(H)$ and $P(T)$ represent prior probabilities and are unconditional, which means that they are not dependent on others. A tuple's posterior probability 'T' is calculated against every class 'C', which means that the classifier is going to indicate the tuple belonging to a class of the maximum posterior likelihood. Then, the equation above is represented more specifically as:

$$P(Ci|T) = \frac{P(T|Ci)P(Ci)}{P(T)} \qquad (8)$$

In the case where the classes C (C1, C2, and Cm) then the classifier finds:

$$P(Ci|T) > P(Cj|T) \quad for \;\; 1 \leq j \leq m, j \neq I$$

$P(Ci|T)$ Is referred to as the maximal posterior hypothesis. Due to the fact that P(T) is constant and the prior probabilities P(Ci) are assume equal, in other wors,

$$P(C1) = P(C2) = \cdots = P(Cm).$$

As a result, $P(T|Ci)$ is maximized.

- *Evaluation Measures*
- Precision of the class represents the likelihood of an arbitrary sentence has to be categorized with this class, in this case it is the precise decision. For example positive class precision is obtained as:

$$P = \frac{TP}{TP + FP} \qquad (9)$$

- Recall of a class represents the possibility that in the case where an arbitrary sentence is to be categorized with the class, then it's the taken decision. A positive class recall is calculated based on the following Equation(10):

$$R = \frac{TP}{TP + FN} \qquad (10)$$

- F-Measure of a class represents harmonic weighted average for each of the calculated recall and precision. F1-measure has been utilized, in order to evenly weigh each of recall and precision. F-Measure is computed according to the following equation [27]:

$$F = \frac{2 * P * R}{P + R} \qquad (11)$$

- *Results and Discussion*

As discussed in earlier sections, an application can be used to social communication between blind and dumb the blind can speak through telephone, which would be recognized by our system. About, 5600 voices samples collected from different speakers are considered for training and testing by using cross-validation. It was observed that the system works efficiently. The Table 1 and Figure 3 illustrate the classification algorithms for the system under various Datasets.

Table 1. shows the classification algorithms

Type	Measures	J48	Naïve byes	KNN
Letters	Precision	100.0	20.012	94.023
	Recall	100.0	20.007	94.164
	F-measure	100.0	18.137	93.510

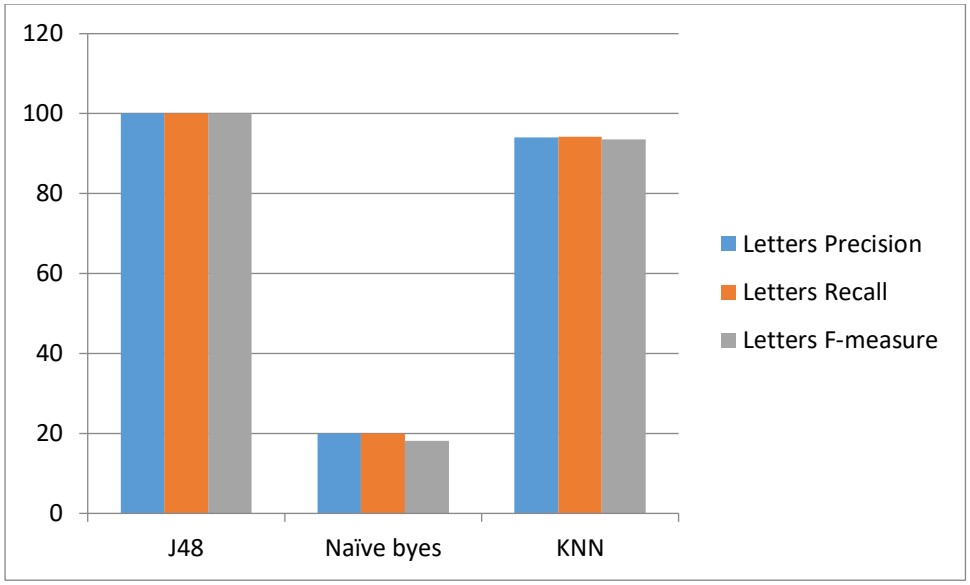


Figure 3. shows the classification algorithms

5. Conclusion

The aim of this paper is to suggest a system for social communication between the blind and the dumb by recording the blind voice and analysing it using MFCC, J48, KNN and Naive byes techniques. Feature extraction was performed using Mel Frequency Cepstral (MFCC) coefficients and classification using J48, KNN, and Naive byes methods. The extracted features are stored in a .WAV file using the MFCC algorithm. Experimental results were analysed with by using Java, and the results were proven to be effective. The experimental shows that J48 is the best nonlinear technology in the ranking with a performance rate of 100 % while KNN 94.023%, Naive byes 20.012% for the letters (أ-ي).

References

[1] Zahoor Mosaad Aydam 2018 *Social Communaction Between the Dumb and Blind*(the College of Education for Pure Science, University of Thi-Qar).

[2] Zulfiqar Ibrahim Merhej and Ghadeer Ahmed Hamed 2016 *Isolated Word Recognition* (Tishreen University Faculty of Information Engineering Department of Software and Information Systems).

[3] Anjali Bala, Abhijeet Kumar And Nidhika Birla 2010 *Voice Command Recognition System Based on MFCC And DWT*(International Journal of Engineering Science and Technology Vol. 2 N0 12).

[4] Eng. Rama hasan 2017 *Improve the Results of the voice recognition based on the result of the integration of different systems*(Tishreen University Electrical and Mechanical Engineering Department of computer and Automatic control Engineering)

[5] Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar 2010 *A Review on Speech Recognition Technique*(International Journal of Computer Applications Vol 10, No.3).

[6] Nidhi Desai , Prof.Kinnal Dhameliya and Prof.Vijayendra Desai 2013 *Feature Extraction and Classification Techniques for Speech Recognition*(International Journal of Emerging Technology and Advanced Engineering Vol 3, Issue 12).

[7] K. M. Shiva Prasad, G. N. Kodanda Ramaiah and M. B. Manjunatha 2017 *Speech Features Extraction Techniques for Robust Emotional Speech Analysis/Recognition*(Indian Journal of Science and Technology Vol 10, No 3) .

[8] Swathy M S and Mahesh K R 2017 *Review on Feature Extraction and Classification Techniques in Speaker Recognition*(International Journal of Engineering Research and General Science Vol 5, Issue 2).

[9] Hernan Faustino Chacca Chuctaya, Rolfy Nixon Montufar Mercado and Jeyson Jesus Gonzales

Gaona2018 *Isolated Automatic Speech Recognition of Quechua Numbers using MFCC, DTW and KNN*(International Journal of Advanced Computer Science and Applications Vol 9, No 10)

[10] Om Prakash Prabhakar and Navneet Kumar Sahu 2013 *A Survey On: Voice Command Recognition Technique*(International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 5).

[11] Dr. Jaffar Alkhier 2017 *Improvement of Speech Recognition by Merging Two Features Extraction Algorithms*(Tishreen University Journal for Research and scientific- Engineering Sciences Vol 39, No 1).

[12] Sheya Narang and Ms.Divya Gupta 2015 *Speech Feature Extraction Techniques*(International Journal of Computer Science and Mobile Computing Vol 4 ,Issue.3)

[13] Shivanker Dev Dhingra, Geeta Nijhawan and Poonam Pandit 2013 *ISOLATED SPEECH RECOGNITION USING MFCC AND DTW*(International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 8).

[14] Namrata Dave 2013 *Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition*(International Journal for Advance Research In Engineering and Technology, Vol 1, Issue V1).

[15] Parwinder Pal Singh and Pushpa Rani 2014 *An Approach to Extract Feature using MFCC*(IOSR Journal of Engineering Vol 04, Issue 08).

[16] Koustav Chakraborty, Asmita Talele and Prof. Savitha Upadhya 2014 *Voice Recognition Using MFCC Algorithm*(International Journal of Innovative Research in Advanced Engineering, Vol 1, Issue 10).

[17] Ms.Rupali S Chavan and Dr. Ganesh. S Sable 2013 *An Overview of Speech Recognition Using HMM*(International Journal of Computer Science and Mobile Computing Vol 2,Issue 6)

[18] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi 2010 *Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*(Journal Of Computing Vol 2, Issue 3).

[19] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad and Arpit Bansal 2013 *Feature Extraction using MFCC*(International Journal (SIPIJ) Vol 4, No 4).

[20] Tina R. Patil and Mrs. S. S. Sherekar 2013 *Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*(International Journal Of Computer Science And Applications Vol 6, No 2).

[21] Nisha 2017 *Voice Recognition Technique*(International Journal for Research in Applied Science & Engineering Technology (IJRASET) Vol 5, Issue V)

[22] Anshul Goyal and Rajni Mehta 2012 *Performance Comparison of Naïve Bayes and J48 Classification Algorithms*(International Journal of Applied Engineering Research Vol 7 No 11),

[23] Uzair Bashir and Manzoor Chachoo 2017 *Performance Evaluation of J48 and Bayes Algorithms for Intrusion Detection System*(International Journal of Network Security & Its Applications (IJNSA) Vol 9, No 4).

[24] Susan Dumais, John Platt and David Heckerman *1997 Inductive Learning Algorithms and Representations for Text Categorization*(Mehran Sahami Computer Science Department Stanford University*).*

[25] S B Harisha, S Amarappa and S V Sathyanarayana 2017 *Kannada Speech Recognition Using MFCC and KNN Classifier for Banking Applications*(International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 1).

[26] Tb. Ai Munandar and Edi Winarko 2015 *Regional Development Classification Model using Decision Tree Approach* (International Journal of Computer Applications Vol 114 , No 8)

[27] Vincent Labatut and Hocine Cherifi 2002 *Accuracy Measures for the Comparison of Classifiers*(Remote Sens Environ vol 80)

Feature Extraction Methods: A Review

Wamidh K. Mutlag[1], Shaker K. Ali[2], Zahoor M. Aydam[3] and Bahaa H.Taher[4]

1Al shatrah Technical Institute, Southern Technical University, Iraq.
2,3Computer sciences and mathematics college, University of Thi_Qar, Thi_Qar, Iraq.
4Information Technology college, University of Sumer, Thi_Qar, Iraq.
1wamid.almuhaysen@stu.edu.iq, 2 shaker@utq.edu.iq ,3 zahoormosad@gmail.com and 4ghrabiuk@gmail.com

Abstract..Feature extraction is the main core in diagnosis, classification, clustering, recognition ,and detection. Many researchers may by interesting in choosing suitable features that used in the applications. In this paper, the most important features methods are collected, and explained each one. The features in this paper are divided into four groups; Geometric features, Statistical features, Texture features ,and Color features. It explains the methodology of each method, its equations, and application .In this paper, we made acomparison among them by using two types of image ,one type for face images (163 images divided into 113 for training and 50 for testing) and the other for plant images(130 images divided into 100 for training and 30 for testing) to test the features in geometric and textures. Each type of image group shows that each type of images may be used suitable features may differ from other types.

Keywords: Geometric features, statistical features, Textures features and Color features.

1.Introduction

In image processing technology, whether it is binary, colored or gray. Image processing may be performed by extracting features for identification, classification, diagnosis, classification, clustering, recognition and detection. Feature extraction method are utilized to obtain much information as possible of image. The selection and effectiveness of feature chosen and extraction are a major challenge now[1]. Many methods used to extract features, which may depend on Geometric features, Statistical features, Texture features, and Color features. Each main type of feature divided into many subdivided types such as Color features divided into three types (Color moment, Color histogram and Average RGB)[2]. Figure 1 shows the most important features methods.

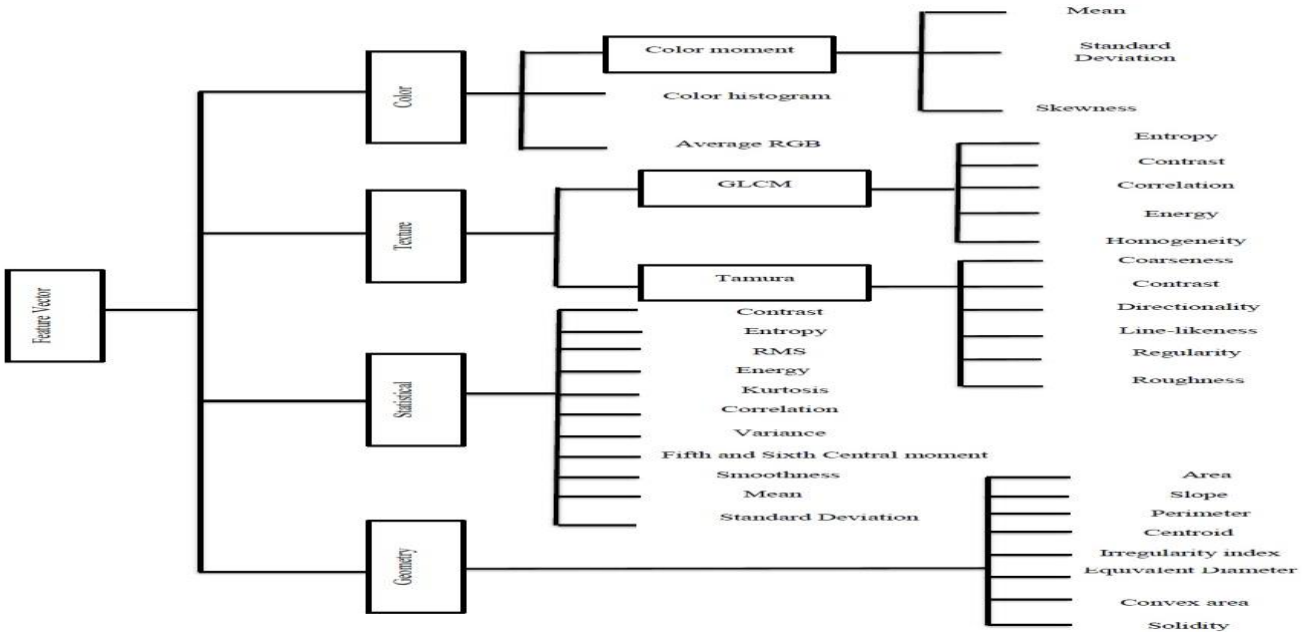


Figure 1. Feature Extraction Methods

2.Feature Extraction

The features can be divided into four types; Geometric features, Statistical features, Texture features ,and Color features. Where each type are further divided into sub-types as following:

1.1. Color Features

The color features are divided into three types (Color moment, Color histogram and Average RGB) [3]

1.1.1. Color Moments:

It can be define as scales that can differentiate images based on their own color feature. The color Moments in the image interpreted as the probability distribution. The three-color moments (Mean, Standard Deviation and Skewness).

1.1.1.1. Mean

The mean can be defined as the mean color value in the image, which mav define in Equation 1

$$M_j = \sum_{m}^{i=1} \frac{1}{M} P_{ji} \quad (1)$$

1.1.1.2. Standard Deviation (STD)

The standard deviation is the square root of the distribution variation, Equation 2 explain the format of STD

$$\mathfrak{6}_j = \sqrt{\frac{1}{M}\sum_{M}^{i=1}(P_{ji} - M_j)^2} \quad (2)$$

1.1.1.3. Skewness

Interpret the deviation as a measure of the degree of asymmetry in the distribution [2]

$$S_j = \sqrt[3]{\frac{1}{M}\sum_{M}^{i=1}\left(P_{ji} - M_j\right)^3} \quad (3)$$

1.1.2. Color Histogram

Color is the most common characteristic and widely used because of its intuitive compared to other qualities and more important information, the ease of extraction from the image and the histogram distributes colors using a set of boxes.

1.1.3. *Average RGB*

The goal of using this feature is for image filtering when using various features. The second reason for choosing this feature because using a small number of data to represent vector parameters [4].

2.2.Texture Features

Texture is the most important feature for many types of images that appear everywhere in nature such as medical images and sensor images and so on. The texture defined as a superficial manifestation of the human visual systems of natural objects. It is easy to recognize by everyone, but it is difficult to determine the texture in matrix, but it occurs in an area of matrix that analyzed texture by quantitative and qualitative analysis. In this paper, two types of texture features are discussed (GLCM and Tamura)[5].

2.2.1.Gray Level Co-occurrence Matrices (GLCM)

A histogram to measure gray values that occur at a given offset on an image. Used to extract texture from a broken tissue image [6]. Those are five various texture features specific by GLCM Entropy, Contrast, Correlation, Energy and Homogeneity [7].

2.2.2.1 Entropy

A statistical measure of randomness be utilized to distinguish the texture of an input image.

$$\text{Entropy} = -\sum\sum \text{q(i, j)}\log\text{q(i, j)} \quad (4)$$

Where q is the number of gray-level co-occurrence matrices in GLCM

2.2.2.2 Contrast:

Calculates the density contrast between pixels and adjacent pixels to whole image. Equation 5 explain the contrast.

$$\text{contrast} = \sum (\text{i,j})^2 \text{q(i,j)} \quad (5)$$

Where, q(i,j) = pixel at location (i,j).

2.2.2.3 Correlation

The function of this scale is to measure the probability of the specified of the specified pixel pairs as in Equation 6 [11].

$$correlation = \frac{\sum_{i=o}^{M-1}\sum_{j=0}^{M-1}(i-n_i)(j-n_j)q(i,j)}{\sigma_i\sigma_j} \quad (6)$$

2.2.2.4 Energy:

Is the summation of squared elements in the GLCM .It is also known as the angular second moment or uniformity.

$$\text{Energy} = \sum\sum \text{q(i,j)}^2 \quad (7)$$

2.2.2.5 Homogeneity

It used to measure the approximation of the distribution of elements in the GLCM to the GLCM diagonal, which define in Equation 8.

$$\text{Homogeneity} = \sum_{i,j} \frac{q(j,i)}{1+|j-i|} \quad (8)$$

2.2.2.Tamura

A description of the quantitative analysis Tamura provided six properties and gave a common description on all types of images texture. Those six various texture features specific by Tamura Contrast, Directionality, Coarseness, Roughness, Line-Likeness and Regularity [8][5].

2.2.2.1.Coarseness

Essentially connect to the distance in the gray levels of spatial changes, which implicitly linked to the size of primeval elements formation the texture. It's have the directly connection to scale and duplication averages and maximum primary texture feature. An image will include iterative textures pattern at different scales, Coarseness tries to find the largest size in which the tissue is present, even in the case of a smaller tissue, as in Equation 9.

$$R_M(x,y) = \sum_{i=x-2^{M-1}-1}^{x+2^{M-1}} \sum_{j=y-2^{M-1}-1}^{y+2^{M-1}} \frac{F(i,j)}{2^{2M}} \quad (9)$$

Where, $2^M * 2^M$ size is the average of neighborhood.

$$S_{M,h}(x,y) = |R_M(x+2^{M-1},y) - R_M(x-2^{M-1} \quad (10)$$

Equation 10 calculates the difference between pair of averages c nding to non-overlapping neighborhoods.

2.2.2.2.Contrast

Measurement allocation of gray levels that change for extent is its distribution into black or white. Determine the contrast, use the central moments of the fourth order of gray levels and the second order

$$Contrast = {}^{\sigma}/_{\alpha_4}$$
$$\alpha_4 = {}^{N_4}/_{\sigma^4} \quad (11)$$

where, $N_4$is the fourth moment about the mean and 2 is the variance. m=1/4 to give the closest value according to Tamura.

2.2.2.3.Directionality

Directionality of an image is measures by which the frequency of local edges directed against directional angles distributed. It is a global property over a region. This feature cannot distinguish between trends or patterns but measures the overall degree of directivity in an image by directivity .The most important feature

among Tamura features through the matrix is to distinguish between one image and another in the consistency of the region.

$$\text{Directionality} = 1 - rm_{peaks} \sum_{p=1}^{m_{peaks}} \sum_{b \in w_p} (b - b_p)^2 H_{directional} \quad (12)$$

Where

m_p:number of peaks

b_p:is the position of the peak

w_p: is the range of the angles attributed to the Pth peak

r: denotes a normalizing factor related to quantizing levels of the angles b and b denotes quantized directional angle

$H_{\text{directionality}}$: is the histogram of quantized direction values,

b :is constructed by counting number of the edge pixels with the corresponding directional angels.

2.2.2.4.Line-Likeness

Line-likeness refers only the shape of the texture primitives. A line-like texture has straight or wave like primitives whose orientation may not be fixed. Often the line-like texture is simultaneously directional. Line-likeness (flin) can be computed as follows

$$flin = \Sigma\Sigma PDd(i,j)njmi\cos[(i-j)2\pi n]\Sigma\Sigma PDd(i,j)njmi \quad (13)$$

Where PDd(i, j) is n×n local direction co-occurrence matrix of points at a distance d.

2.2.2.5.Regularity

Regularity measures a constant pattern or comparable in an image. define in Equation 14.

$$Y_{\text{Regularity}} = 1 - R(C_{CRS} + C_{con} + C_{dir} + C_{lin}) \quad (14)$$

where R is a normalizing factor and Cxxx means the standard deviation of fxxx

2.2.2.6.Roughness

Is the sum of the measures of coarseness and contrast

$$Roughness = Coarseness + contrast \quad (15)$$

2.3.Statistical Features

Approaches do not attempt to understand explicitly the hierarchical structure of the texture. Instead, they represent the texture indirectly by the non-deterministic properties that govern the distributions and relationships between the grey levels of an image. eleven types of texture features are discussed (Contrast , Entropy , RMS , Energy, Kurtosis, Correlation, Variance, Fifth and sixth central moment, Smoothness, Mean and standard deviation)[9-10].

2.3.1.Contrast

Contrast is a measure of intensity or gray-level variations between the reference pixel and its neighbor. Contrast is determined by the brightness of the object color, and other objects within the same display area. define in Equation 16

$$F = \sum_{m}^{M_g-1} m^2 \left[\sum_{i=0}^{M_g-1} \sum_{j=0}^{M_g-1} q_{d,\theta}(i,j)\right] \quad (16)$$

Where m=|i-j| When i and j are equal, the cell is on the diagonal and i − j = 0. These values appear pixels entirely comparable to their neighbor, so there are specific a weight of 0. If I and j differ by 1, there is a small contrast, and the weight is 1. If i and j differ by 2, the contrast is increasing and the weight is 4. The weights continue to increase exponentially as (i − j) increases.

2.3.2.Entropy

It is used to measure the system disturbance in the physics of thermodynamics. Entropy measurement is an ideal way to measure the level of unstable signal disturbance as well as to measure the amount of information contained in the event. define in Equation 17

$$\text{Entropy} = -\text{sum}\left(P * \log(P)\right) \quad (17)$$

Where: P:probability vector.

2.3.3.RMS(Root mean square error)

The RMS value Gradually increase the value with the development of error However, the unable to provide the information Special of incipient fault stage while The value increases gradually as the error develops as define in Equation 18

$$R = \sqrt{\frac{1}{M}\sum_{j=1}^{M} |y_j|^2} \quad (18)$$

2.3.4. Energy

It is utilize to describe a measurement of information while Perform an operation under a probability frame this as (maximum a priori) assessment in coupling with Markov Random domain. Sometimes the energy can be a positive measure to maximize and sometimes it is a negative measure to minimize. It is specific by mean of [19]

$$F = \sum_{j}\sum_{i} q(j,i)^2 \quad (19)$$

2.3.5.Kurtosis

It Measures the stability of the distribution, which relates to the normal distribution

$$\text{Kurtosis} = \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{(q(j,i)-m)^4}{(MN)\sigma^4} \quad (20)$$

2.3.6.Correlation

Correlation is the basic process used to extract the information from the image as shown in the following Equation 21

$$\text{Correlation} = \frac{\sum_i \sum_j q(i,j) - M_x M_y}{\sigma_x \sigma_y} \quad (21)$$

2.3.7.Variance

The variance is define as the mean of the signal square and is calculated after the mean value is removed, define in Equation 22

$$\sigma^2 = \frac{1}{q}\sum_{i=1}^{q}(Y_j - M)^2 \quad (22)$$

Where: σ =Variance,q=no of samples, Y_j=input heart signal μ= mean

2.3.8. Fifth and sixth central moment:

That give the deviation about average. Fifth central moment,

$$= \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{(q(j,i) - m)^5}{(MN)\sigma^5} \quad (23)$$

Sixth central moment

$$= \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{(q(j,i)-m)^6}{(MN)\sigma^6} \quad (24)$$

2.3.9. Smoothness

Comparative smoothness, Q is a measurement of gray level disparity that which can used to create relative smoothness recipes. The smoothness is specific by Equation 25

$$Q = 1 - \frac{1}{1+\sigma 2} \quad (25)$$

Where, σ is the standard deviation of the image

2.3.10. Mean

Calculates the average values in the image

$$\text{Mean} = \sum_{i=1}^{r}\sum_{j=1}^{t} \frac{q(i,j)}{rt} \quad (26)$$

Where q(i, j) is the intensity value of the pixel at the point (i,j). The image is of r by t size.

2.3.11.standard deviation

Calculates the mean distance between the pixel value and the mean where the low standard deviation value indicates that there is less deviation of the pixels from the mean and the higher value indicates the high contrast, define in Equation 27

$$\sigma = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{t}\frac{(q(i,j)-m)^2}{rt}} \qquad (27)$$

2.4.Geometry Features

There are eight types of geometry features as following :

2.4.1. Area

Is the extension of shapes, and it is different from the perimeter. Where the linked inside the shape, there are many known formulas for simple forms such as triangles, rectangles, and circles. Using these formulas, any polygon area can be calculated by dividing the polygon into triangles or circles to obtain curved shapes with borders and then collected after the calculation of their areas and when the polygon is irregular can polygon area is calculated by equation Gauss trapezoidal and described as in the following Equation(28)[11]

$$A = \frac{1}{2}\sum_{i=0}^{m-1}(p_i * q_{i+1}) - (p_{i+1} * q_i) \qquad (28)$$

Where: m: number of points , p_i : X axis coordinates, q_i : y axis coordinates

2.4.2.Slope

The straight line is a set of points, that which has a fixed slope between any two points. The slope of the straight line is usually determine by the value of the ratio of vertical change to horizontal variation. The slope usually describes the slope of the two-point line. The parallel line of the x-axis is define as the horizontal line, Zero. The parallel line of the y-axis known as the vertical line, and its slope always has an undefined value. The parallel two lines always have slope equal [12].This is described by the following Equation (29)

$$\text{Slope} = \frac{pz-po}{qz-qo} \qquad (29)$$

2.4.3.Perimeter

It is the length of the line that surrounds of two-dimensional shapes such as; the circle, square, rectangle or irregular shapes. The perimeter can be calculated as in Equation (30) if the shape is equilateral while the equation (31) calculates the perimeter if the shape is ribbing inequilaterally [13].

$$\text{Per} = n * (x) \qquad (30)$$

$$\text{Per} = \sum_{i=0}^{n-1} x_i \qquad (31)$$

Where: n is number of ribs, x:length of the rib

2.4.4. Centroid

The centroid is a fixed point in the object where the lines pass through this point, which represents the weight of the object. The centroid is different from each other in terms of form or acclimatization and thus determine the status of a centroid related to this difference. The centroid can calculate according to the following Equation (32) [14].

$$x_o = \frac{\sum x_{oi}A_i}{\sum A_i},\ y_o = \frac{\sum y_{oi}A_i}{\sum A_i} \qquad (32)$$

Where: x_o: the x- axis value when center point of shape

y_o: the y- axis value when center point of shape

x_{oi}: The distance at which the center of the shape be far from the junction point of the axes on axis (x)

y_{oi}: The distance at which the center of the shape be far from the junction point of the axes on axis (y)
A_i: area the shape.

2.4.5. Irregularity Index[15]:

The boundaries of irregular shapes is calculated by the Equation(33)

$$L = \frac{4\pi * A}{per} \quad (33)$$

Where: A:is area, Per: is perimeter

The metric value irregularity index (L) is equal to one only for circle and it is < 1 for any other shape.

2.4.6. Equivalent Diameter [16]:

Numerical that determine the diameter of the circle together with the same region as the area. It is calculate as in the following Equation (34).

$$Q_{diameter} = \sqrt{\frac{4*A}{\pi}} \quad (34)$$

Where: A: is area

2.4.7. Convex Area

The closed convex or convex represents the set X of points in the Euclidean level the smallest convex set contains X.For example, when X you are a limited subset of the plane. The convex area is the number of pixels in the convex image.The size of the square surrounding the area.Where the bounding box is a convex hull[17].

2.4.8. Solidity

Calculates the pixel ratio in a convex hull located in the area[18]

$$\text{Solidity} = \frac{A}{\text{Convex Area}} \quad (35)$$

Where: A: is area

3. Comparison and analysis

In this section, discuss the comparative results of features extraction and database used for general purposes including about 280 images by two categories with the KNN algorithm for classification as show in Table.1.

4. Discussion

In this paper, four main types and subtypes for each type of features extraction are collected, each feature extraction is necessary for special application anyway. The similarity measurement of efficiency conclusion includes TP values and TN values [19]. The equation (36) shows the accuracy four main types of extraction features and subtypes

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN} \quad (36)$$

5**. Conclusions**

This section displays the database including two different categories and accuracy are gain from this review.

Table 1 portrays the features and the database used

Database Contents	Number of Samples	Type of features used	Acuracy	
			Training	Testing
Face	Training:113 Face images Testing :50 Face images	Geometry	100%	98%
		Statistic	99.5%	96%
		Color	98.8%	96%
		Texture	97%	95%
Plant	Training:100 Plant images Testing :30 Plant images	Geometry	100%	93%
		Statistic	100%	95%
		Color	100%	97%
		Texture	100%	98%

References

[1] Ethem Alpaydin 2014 *Introduction to Machine Learning* (MIT Press).

[2] Kavya R and Harisha 2015 *Feature Extraction Technique for Robust and Fast Visual Tracking: A Typical Review* (International Journal of Emerging Engineering Research and Technology Vol 3 Issue 1) PP 98-104,

[3] S.R. Kodituwakku and S.Selvarajah 2011 *Comparison of Color Features for Image Retrieval* (Indian Journal of Computer Science and Engineering Vol 1, No 3) pp 207-211.

[4] Amera H.M alzoubi 2015 *Comparative Analysis of Image Search Algorithm using Average RGB, Local Color Histogram Global Color Histogram and Color Moment HSV*(thesis, Faculty of Computer Science and Information Technology Universiti Tun Hussein Onn Malaysia).

[5] C. Umamaheswari, Dr. R. Bhavani and Dr. K. Thirunadana Sikamani 2018 *Texture and Color Feature Extraction from Ceramic Tiles for Various Flaws Detection Classification* (International Journal on Future Revolution in Computer Science & Communication Engineering Vol 4 ,Issue. 1)pp 169 – 179.

[6] M. S. Ahmad, M. S. Naweed, and M. Nisa 2009 *Application of texture analysis in the assessment of chest radiograph* (International Journal of Video & Image Processing and Network Security (IJVIPNS) Vol 9,No 9) pp 291-297.

[7] .Fritz Albregtsen 2008 *Statistical Texture Measures Computed from Gray Level Co-occurrence Matrices* (Image Processing Laboratory Department of Informatics University of Oslo).

[8] Peter Howarth and Stefan Ruger 2004 *Evaluation of Texture Features for Content-Based Image Retrieval* (Department of Computing, Imperial College London South Kensington Campus London).

[9] Monika Sharma, R. B. Dubey and Sujata and S. K. Gupta *2012 Feature Extraction of Mammograms*(International Journal of Advanced Computer Research Vol.2, No.3)

[10] Jaspinder Kaur, Nidhi Garg and Daljeet Kaur 2014 *Segmentation and Feature Extraction of Lung Region for the Early Detection of Lung Tumor*(International Journal of Science and Research (IJSR) Vol 3, Issue 6).

[11] A. H. Stroud 1971 *Approximate Calculation of Multiple Integrals*(Prentice-Hall Inc., Englewood Cli_s, N. J.).

[12] Christopher Clapham, James Nicholson 2009 *Oxford Concise Dictionary of Mathematics*(OUP oxford).

[13] Dr Yeap Ban Har,Dr Joseph Yeo,Teh Keng Seng,Loh Cheng Yee,Ivy Chow,Neo Chai Meng and Jacinth Liew 2018 *New Syllabus Mathematics Teacher's Resource Book1*(OXFORD UNIVERSITY Press) 7th edition.

[14] Dan B. Marghitu and Mihai Dupac 2012 *Advanced Dynamics* (Springer) chapter 2 pp 73-141.

[15] S. A. Patil and V. R. Udpi 2010 Chest x-ray features extraction for lung cancer classification(Journal of Scientific and Industrial Research Vol 69) pp 271-277.

[16] R. C. Gozalez and R. E. Woods 2002 *Digital Image Processing Using Matlab*, 2nd ed, Gatesmark (USA) chapter 12 pp 642-654.

[17] Nitin S. Lingayat and Manoj R. Tarambale 2013 *A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image*(International Journal of Bioscience, Biochemistry and Bioinformatics Vol 3, No. 6) .

[18] K. P. Aarthy and U. S. Ragupathy 2012 *Detection of lung nodule using multiscale wavelets and support vector machine*(International Journal of Soft Computing and Engineering (IJSCE) Vol 2, Issue 3).

[19] Mariam A.Sheha , Mai S.Mabrouk and Amr Sharawy 2012 *Automatic Detection of Melanoma Skin Cancer using Texture Analysis*(International Journal of Computer Applications Vol 42,No 20)

Review: A comparison Steganography Between Texts and Images

Assist. Prof. Dr. Maisa'a Abid Ali Khodher **** Assist. Lec. Teaba Wala Aldeen Khairi**

Department Computer Sciences/ University of Technology-Iraq
*110044@uotechnology.edu.iq *110053@uotechnology.edu.iq

Abstract: The steganography is branch from information hiding, the speed of evolve communications Internet and networks in wide areas in the world. This evolve make to most people tends for work in security data through transmit across networks from sender to receiver. The major aim from steganography uses to protect the substantial data, such as text, image, video, and audio during transmit between sender and receiver.

The problems in steganography, because the people to increase uses internet in the present time therefore, needs to protected information during transmitted from sender to receiver. And solve this problem in steganography, in here many techniques is used in this article.

This article offers comparing in steganography techniques between texts and Images, when hiding secret message in texts and Images. In this study, several techniques it uses by researcher in domain of steganography.

The outcomes to obtained comparing between steganography texts or images. The results that shown the compression between text and image steganography are good and efficiency together without sensitive by attackers.

1. Introduction

Steganography is one of the most effective secured data communication. It supplies a security to secure letter via embedded them into digital mediums and make them not clear and not visible for eavesdroppers [1]. The dissimilarity to the conventional cipher which objective is to conceal the content of secure letters being interchanged among the two connection parties, the objective of information hiding is to conceal not only secret message but also its not existence. Thus, it can offer a best security in several methods. There are two another technology that are in similar correlation regarding to information hiding; they are watermarking and fingerprinting [2] which include the embedded of data in some mediums.

The information hiding has sciences of hiding secure data at any mediums like picture, sound, video...etc. whereas no eavesdropper can be empathy with secure communication. In the steganography it uses a high security concealing methods using DWT and optimize letters dispersing manner. Sometime it is applied HWT to the covering picture in higher hesitation and lower hesitation data and higher hesitation data includes data on border, angle. etc. for picture which is dispersed our secure data. Security letter was entries in every color component of higher hesitation bands that are Red, Green and Blue color the compounds start for the final column of every at color the compounds where up of down rely on the extent of letter [3].

2. Steganography Type

There are three kinds of key in steganography, they are pure key, secrete key, and public key.

1- Pure Key:

The pure key is using in any media (text, image ...etc.) to hide secret message, in this key no demand before interchange the of some secret data. In this key rely on entirely on its securely [4]. The pure key is defined (C, M, D, E), as shown in Figure 1.

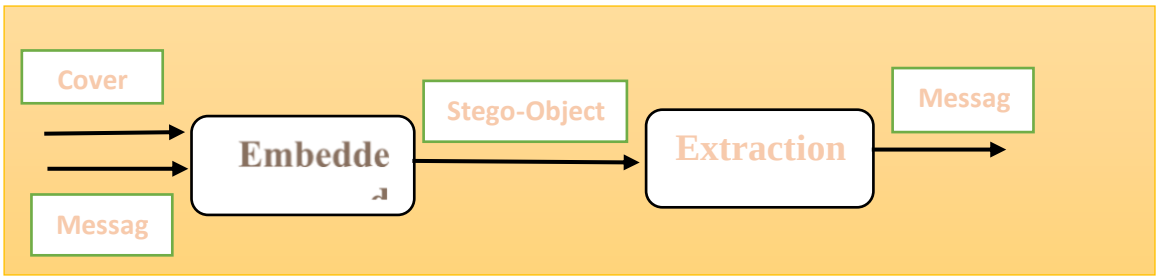


Figure 1: The pure key.

2- Secret Key:

The secure key like symmetrical key in encipher, where sender choose cover to embed the secret message, it uses found location from cover using secret key to hide this secret message. The secret key used embedded process must be known for receiver, it can extraction this secret message [4], [5]. The secret key is defined (C, M, K, DK, EK), as shown in Figure 2.

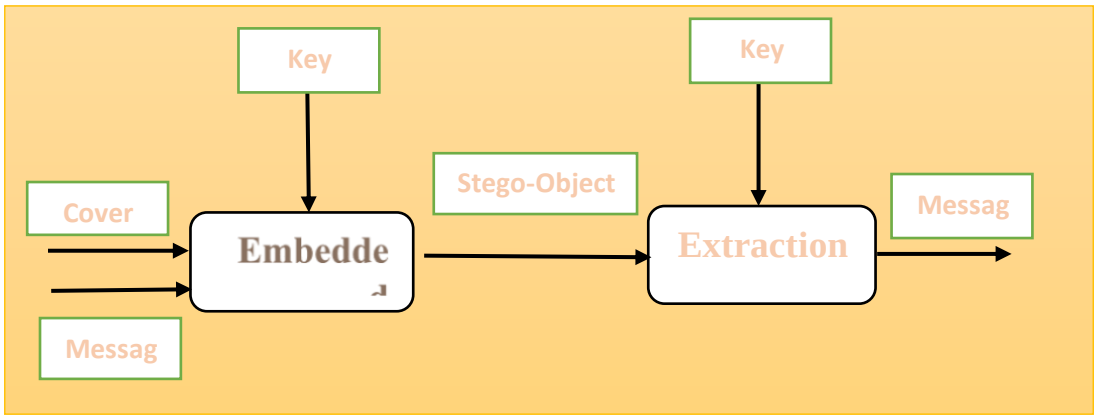


Figure 2: The secret key.

3- Public key:

The public key not rely on interchange for secure key. It is demands pair keys, the private key is the first and the public key is the second. The public key is saved in database, while it is using the public key the embedded operation. The secure key is using to extraction the secure letter [5].

3. Previous Work in Steganography Text

In 2010, Adnan Abdul-Aziz Gutub, et al, proposal an enhanced Arabic script information hiding method to Arabic script it uses kashida. The method conceals secure data as bit into Arabic characters (covering) through utilizing extension letter (kashida). This method is considering, when the secret bit is (0) put one kashida, and when secure bits are (1) put two kashidas after character who can save it. When final character is embedding exactly next the latter bit of secure data, and can be embedded the kashida randoms to the remaining script in orderly that reinforce the secure method. As their method reinforce secret, the Arabic scripts rely on ability and robust for secure telecommunication [6].

In 2013, Ammar Oden, et al, proposal an enhanced Arabic script information hiding method for Arabic script utilizing different at kashida. This method choice one in four screenplays random of conceal secure data embedding such as bits into Arabic characters (covering) during utilizing kashida. This method deem without-point Arabic characters put a kashida when a secure bit is (0), and point Arabic character put kashida when a secure bit is (1) such as first screenplay, and the second screenplay is vice versa. In third screenplay was added kashida after Arabic character when a secret bit is (1) and (0) is otherwise, and in

fourth screenplay is vice versa. This method reinforce security, complication to Arabic script rely on security telecommunication [7].

In 2018, Kemal Tutuncu, and Abdikarim Abi Hassan, proposal is utilizing from email addresses for keys to embedding/into extraction the secure letter to/from email script (covered script). In after choosing the covered script has higher duplication style as regards to the secret letter the space of array was created. The organ of distance array was compressing by next lossless compressing algorithm is written series; the [RLE], ([BWT],[MTF] , [RLE], [ARI]. The following on Latin Square is using for compose stego-key one and where Vigenere encryption is using to excess complication of extracted stego-key one. End stage is choice e-mail addresses through utilizing stego-key one [K one] and stego-key two [K two] to embedding secure letter within forward email platform. This tests of outcomes display that suggest a manner has sensible execution in terms for space, and as well the highest secure of terms of complication [8].

4. Previous Work in Steganography Image

In 2014, A. Gupta, S. Shantaiya, proposal vary filters and algorithm similar reveres filter, wiener filter and an afflicted Lucy-Richardson deconvolution algorithm. Before deconvolution step, our split up the stain picture into sleek partition. through insert vary noises and picture become corrupted parameter scale and extent, the stain picture are then used for picture deblurring. The outcome on filter compare supply the vary parameter which established the picture goodness and better outcome [9].

In 2014, Abbas F. Tukiwala, and Sheshang D. Degadwala**,** Proposal technique summary by joining the feature of cipher and conceal. ciphering using adjust ASCII transformation and Mathematic job include transform the secure letter at unprintable shape of same volume such as main letter at any status. Information hiding is thereafter used multi-level 2-D DWT to embedded that cipher datum inside a covering medium used higher hesitation Coefficients of every distance into every level at 2-D Haar DWT, and conceal it is presence. lastly, Execution may be measures was used statistical parameter, [PSNR], and [MSE]. The outcome of that technique supply every three side of datum hiding such as "capacity, security and robustness"[10].

In 2018, Maisa'a Abid Ali Khodher, Proposal a new algorithm is proposed that enables secret messages to be embedded inside satellite images, wherein images of any size or format can be hidden, using a system's image compression techniques. This operation is executed in three main steps: **first phase**–the original image is converted into a raster image; **second phase**–steganography, in which a binary secret message is hidden inside a raster image, using a 4×4 array as the secret key; and **third phase**–compression of the stego-image raster in L2 and L3 using a 2-D wavelet packet. The outcome is a highly efficient algorithm, which can rapidly conceal information inside transmitted satellite images, thus guarding against revealing information to potential cyber-attackers [11].

5. Steganography Text methods

The steganography consists of several media which are text, image, video, and audio, as following in Figure 3.

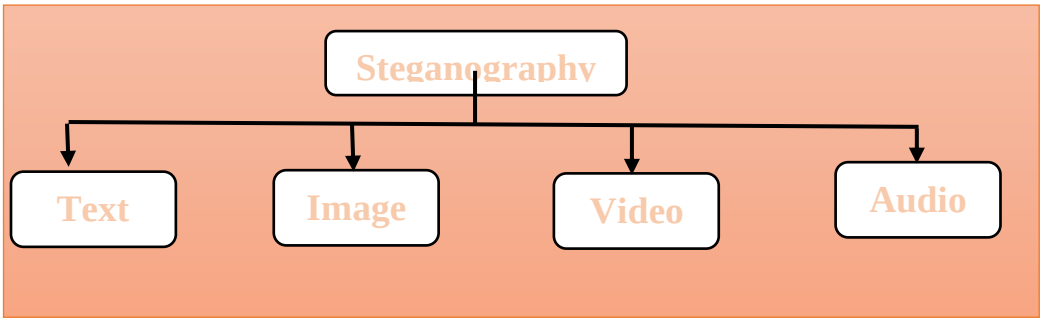


Figure 3: The media of steganography.

This section describes steganography text methods; [Linguistics steganography, Format- based, and Random, and statistical generation]. ***The Linguistic steganography*** include (syntactic, semantic, and lexical). ***The syntactic*** uses point mark (,) or (.), ***the semantic*** uses synonym words, ***the lexical*** uses take list of synonym word using 00, 10, 01, 11 to conceal secret message. ***Random and statistical generation*** (that methods are used to cover-text generated automatic accord to the statistic ownership of language. Because avoid compare within a recognized actual script, steganography oftentimes resorts for generated it's have cover scripts. One manner is data hiding in randomly looking series of letters. Letter series manner conceals the data into letter series) [12], [13].

Format based include (line shift, word shift, whit space, and feature coding). *The* ***line shift*** uses When it hides zero bit, a line is shifted up and when it hides one bit, the line is shifted down, ***word shift*** uses secure letter is conceal the words shifting via horizontal, i.e. left or right to represents zero bit or one bit separate, ***white space*** uses whereas statement spacing is enter, when site single space into hide zero bit and two spaces to conceal one bit at the end of each ending letter, and ***feature coding*** uses dot in message i and j are not accepted, extent of strike in messages f and t can be change, or by extension or lessen rise of messages b, d, h [12]. As shown in Figure 4.

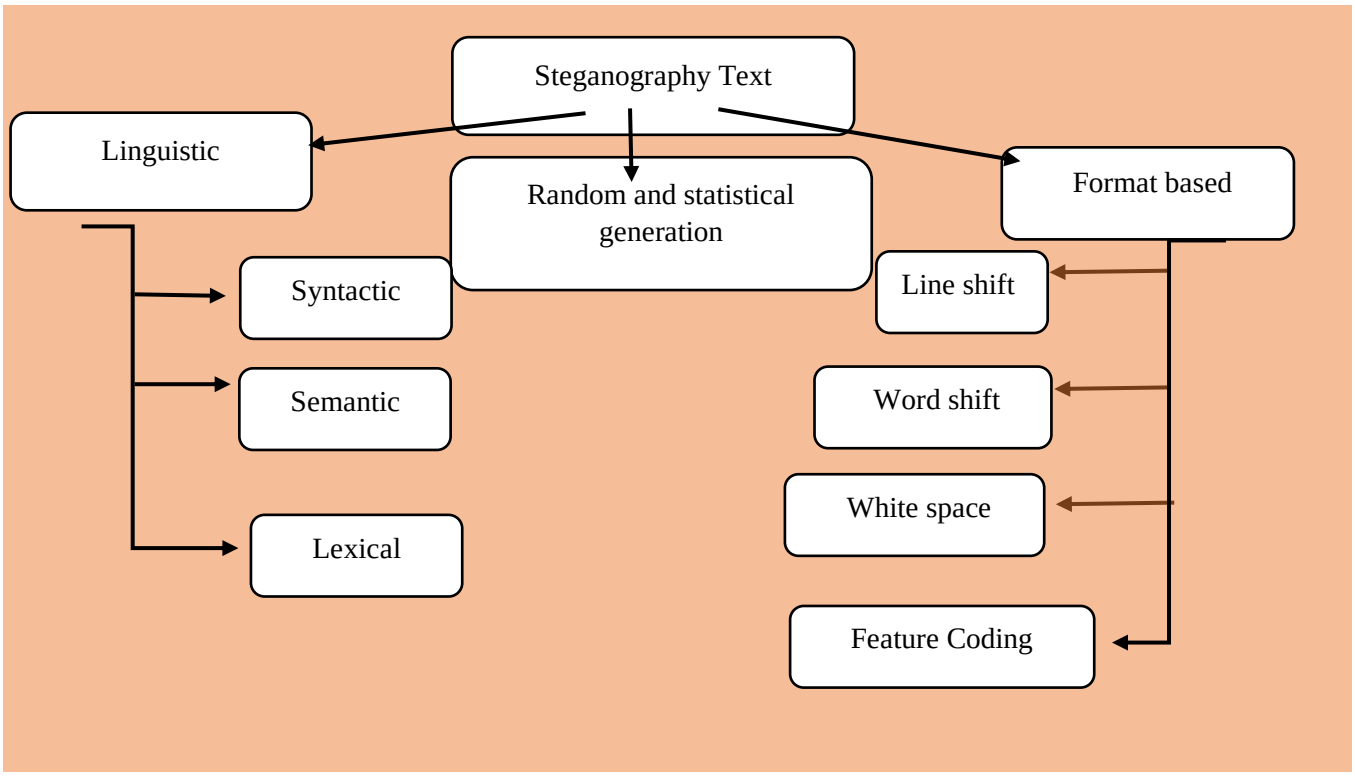


Figure 4: The steganography text methods [4].

6. Steganography image methods

This section offers steganography image methods, spatial domain, transform domain, spread spectrum, statistical, and distortion methods. The spatial domain includes (LSB, PVD, and substitutions), Transform includes (IWT, WDT, and DCT) [14]. As shown in Figure 5.

The (LSB) Least Significant Bit information hiding is easy method of embedded datum into images. LSB method directed embedded the secure datum inside the LSB at the pixel [15], as shown in Figure 6.

Original pixel Secrete data Stego pixel

Figure 6: The LSB techniques.

The Pixel Value Difference PVD can embedding larger amount of information without many dissolutions at the picture quality and so are seldom sensitive through human eyes. PVD is used the vary at every two pixels for determine several of letter bits, when it can be embedding inside the two pixels. It begins of the top-left angle at the covering picture and scans the picture zigzag, as shown in Figure 7 [16].

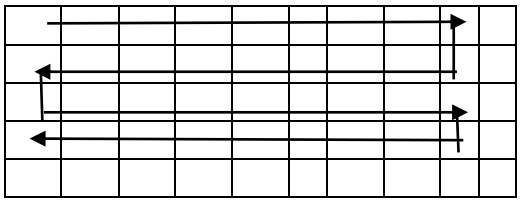


Figure 7: The PVD is zigzag in image.

The (IWT) Integer Wavelet Transform is a type of wavelet transform who maps integer set of datum, with other integer set of datum. IWT have the significant ownership that it is coefficients has the same dynamic domain as the main signals. That make to easy execution consider regard the volume of the variables, it is using and the domains to supply in coding algorithm. It takes four bands convergent, Vertical, Horizontal, and diagonal Bands that appear as LL, LH, HL and HH restively [17], [18]. As shown in Figure 8.

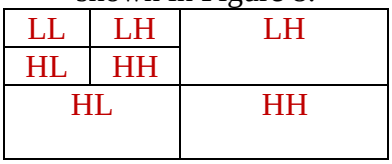


Figure 8: 2 Level Integer Wavelet Transform.

Discrete Wavelet Transform (DWT) is frequency domain usually done using Wavelet transform. The use of wavelets in the form of shorthand model lies in a statement that the wavelet transform is obviously splits the higher from the low- hesitation information based on pixels [18]. The simplest method of wavelet transforms is the Haar wavelet. In Hara, transform the coefficient in the low frequency wavelet was created by take the averaging of the values of two pixels, and it have created the high frequency [17], [18].

The (DCT) Discrete Cosine Transform have been an international standard in Joint Photographic Experts Group (JPEG) form to decrease the blocking impact of image compression. The FDCT algorithm that utilizes the energy compactness and matrix sparseness properties in hesitation area to achieve higher calculated performance. For a JPEG image of 8 × 8 block volume in spatial area, the algorithm decomposed the two-dimension(2D) DCT into one pair of one-dimensional (1D) DCTs within transform. The 2D spatial datum is a linear merge of the rule image obtain during the outer results of the column and vectors of cosine functions so that reverse DCT is as active [19].

7. Comparing between Text and Image steganography

The comparing between text and image in steganography rely on two basic algorithms of embedding and extraction algorithm. The Table 1. Indicates for comparing text and image steganography.

Table 1: comparing between text and image steganography.

Text method	Linguistic	Format based	Capacity	Robustness	Precision
Syntactic	Yes	No	High	High	High
Semantic	Yes	No	Medium	High	High
Lexical	Yes	No	Medium	High	High
Line shift	Yes	Yes	Medium	Medium	High
Word shift	Yes	Yes	Medium	Medium	High
White space	Yes	Yes	High	High	High
Feature coding	Yes	Yes	Low	High	High
Random and statistical generation	yes	No	High	High	High
Image method	**Spatial domain**	**Transform domain**	**Capacity**	**Robustness**	**Precision**
LSB	yes	No	High	Medium	High
PVD	Yes	No	High	High	High
Substitution	Yes	No	Medium	Medium	High
IWT	No	Yes	High	High	High
DWT	No	Yes	High	High	High
DCT	No	Yes	Medium	High	High
Another method	No	No	Medium	Low	Low

Conclusion: This paper offers comparing between texts and images in steganography, the steganography text is very difficult to hide secret message, because text can be visible by human eye, but the steganography image is easy to hide secret message, because the image minor details cannot be visible by human eye.
In steganography text operation in English language more simple than Arabic language, because in Arabic existence movements in characters and sentences and put kashida after or before character, that make to Arabic language more difficult in hiding. Therefore, it can be find methods not effect in texts during hiding secret message. The efficiency, robustness, and high security is very important in this methods.
In steganography image can hide secret message high capacity in spatial and transform domains, because rely on size of image. Image is more efficient for hosting, robustness, and high capacity and high security in hiding secret message.
generally, the size of text smaller than image that indicates capacity of text is least than image, therefore, the size of image takes larger secret message than text.

References

[1] Tian, Lei; Zhou, Ke; Jiang, Hong; Liu, Jin; Huang, Yongfeng; and Feng, Dan, An M-Sequence Based Steganography Model for Voice over IP (2008). *CSE Technical reports.* Paper 68.

[2] Cheddad, Abbas, Joan Condell, Kevin Curran, and Paul Mc Kevitt., Digital image steganography: Survey and analysis of current methods, *Signal Processing* 90, no. 3 (2010): 727-752.

[3] Juned Ahmed Mazumder, and Kattamanchi Hemachandran, Color Image Steganography Using Discrete Wavelet Transformation and Optimized Message Distribution Method,International Journal of Computer Sciences and Engineering (IJCSE), Vol.2, Issue.7, 2014.

[4] Zaidoon Kh. AL-Ani, A.A.Zaidan, B.B.Zaidan, and Hamdan.O.Alanazi Overview: Main Fundamentals for Steganography, Journal of Computering, Vol. 2, No. 3, March 2010: 158-165

[5] Maisa'a Abid Ali K., A Framework to Design and Implementation of A Linguistic Steganography System, Ph.D., University of Technology, 2016.

[6] Adnan Abdul-Aziz Gutub, Wael Al-Alwani, and Abdulelah Bin Mahfoodh, Improved Method of Arabic Text Steganography Using the Extension 'Kashida' Character, Bahria University Journal of Information & Communication Technology Vol. 3, Issue 1, December 2010.

[7] A. Odeh, K. Elleithy, and M. Faezipour, Steganography in Arabic Text Using Kashida Variation Algorithm (KVA), Systems, Applications and Technology Conference (LISAT), 2013 IEEE Long Island, 2013, pp. 1-6.
[8] Kemal Tutuncu, and Abdikarim Abi Hassan, New Approach in E-mail Based Text Steganography, International Journal of Intelligent Systems and Applications in Engineering, IJISAE, Vol. 3, No. 2, 2015, 54-56.
[9] A. Gupta, and S. Shantaiya, Reduction of Image Blurring with Digital Filters, Journal of Engineering Research and Applications, Vol. 4, No.1, 2014, 139-143.
[10] Abbas F. Tukiwala, and Sheshang D. Degadwala, Data Hiding in Image using Multilevel 2- D DWT and ASCII Conversion and Cyclic Mathematical Function based Cryptography, International Journal of Computer Applications (0975 – 8887), Vol. 105 – No. 7, November 2014.
[11] Maisa'a Abid Ali Khodher, Hide Secret Messages in Raster Images for Transmission to Satellites using a 2-D Wavelet Packet, Iraqi Journal of Science, Vol. 59, No.2B, 2018, 922- 933. DOI:10.24996/ijs.2018.59.2B.14.
[12] Xiaoxi Hu, Gang Luo, Yongjing Lu, and Lingyun Xiang, *A Steganography on Synonym Frequency Distribution*, Advances in information Sciences and Service Sciences(AISS), Vol.5, No. 10, May 2013.
[13] M. Agarwal, *Text Steganographic Approches: A Comparison*, International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.1, January 2013.
[14] Nagham Hamid, Abis Aahya, R. Badlisha Ahmad, and Osamah Al-qureshi, *Image Steganography Techniques: An Overview,* International Journal of Computer Science and Security (IJCSS), Vol. 6, No. 3, 2012, 168-187.
[15] Aditya Kumar Sahu, and Monalisa Sahu, *Digital Image Steganography Techniques In Spatial Domain: A Study*, International Journal of Pharmacy and Technology, Vol. 8, No. 4, January 2017, 5205-5217.
[16] El-Sayed M. El-Alfy, and Azzat A. Al-Sadi, *Pixel-Value Differencing Steganography:Attacks and Improvements,* ICCIT, 2012.
[17] Hemalatha S., U Dinesh Acharya, Renuka A., and Priya R. Kamath, *A Secure Color Image Steganography In transform Domain,* International Journal on Cryptography and Inform- ation Security (IJCIS), Vol.3, No.1, March 2013, 17-24.
[18] Iman I. Hamid, *Image Steganography Based on Discrete Wavelet Transform and Chaotic Map,* International Journal of Science and Research (IJSR), Vol. 7, No. 1, January 2018, 588-591.
[19] S. E. Tsai, and S.M. Yang, *A Fast DCT Algorithm for Watermarking in Digital Signal Processor,* Mathematical Problems in Engineering, Vol. 2017, 1-7. https://doi.org/10.1155/2017/7401845

Convert Gestures of Arabic Words into Voice

Shaker K .Ali[1], Ali Al-Sherbaz[2], Zahoor M. Aydam[3]

1,3 Computer Department, Computer sciences and mathematics college, University of Thi_Qar ,Thi_Qar, Iraq
2 Computer Department, Faculty of Science and Technology, University of Northampton, UK

Abstract. Gestures are one of the best ways of communication between dumbs and other people using the expression of signs language. In this paper, we suggest an algorithm for recognizing hand gestures of Arabic words (اتمنى لك حياة سعيدة-اقتباس) to by using dumb (through signs) and convert the sings into voice corresponding to sings words. The proposed algorithm for Convert Gestures of Arabic Words into Voice , record video of gesture (of the dumb person) then convert the video into frames (images), preprocessing for the resulted image must done by remove the noise, resize the images and increase the contrast, then calculate the distance to clustering the words by using (C4.5 , k-mean , k- medoid and artificial neural network), calculate the distance (or features) by using Euclidean distance and slope where ,there are eighteen features (eight features from Euclidean distance, eight features from slop, Area, and perimeter). The results in the training stage were (C4.5 gave 100%, k-mean gave 95.2% k-medoid gave 91.9% and ANN gave 91.27%). While in the testing stage we used three classifiers (Euclidian Distance, Modify of the Standardize Euclidian Distance and Correlation) and the results show that (Euclidian Distance gave 94.4%,Modify of the Standardize Euclidian Distance gave 100% and Correlation gave 94.4%) We create our database (three videos with 250 frames) for training and one video for testing.
Keywords: Gestures, Feature Extraction, C4.5, K-Mean , K-Medoid and ANN

Introduction

Communication is the way for expression about thoughts, opinions, information, or messages between the people by writing, speaking, or signs. Communication is usually oral expression between people by talking to each other while people dumb cannot communicate with others as ordinary people do, they can't speaking people who are deaf are able to speak, but they unable to hear. While the blind are unable to see but they can speaking and listen [1]. The gesture is a kind of nonverbal communication with a part of the body, which used together with verbal communication. The gestures are obscuring not totally specific. Like the talk and handwriting, gestures change from individual to individual, even to the same person in different cases [2].A gesture is a language used by dumb people. Dumb people use signs to show their ideas. Gesture language is different from each country to another country with its special vocabulary and grammarian. In fact, gesture language can vary in one country from one place to another, as Languages spoken [3]. The gesture is the movement of any part of the body such as the face and hands a kind of motion [4]. There are two methods for recognizing the gesture; the first way is based glove and the second was based on computer. The first way depends on the hardware and gets information from the joints of the hand by using sensors to know the classification of hand gesture. This way use video and convert the video into frames to identify the pattern they know the hand gestures [5].

Recognize of gesture language at present, by the token gesture of humans using video camera such as a mobile, tablet, special camera, or laptop camera [6] then convert the video into the image and extract the features then classify each number into voice, this paper focuses on the how the gesture language translate into voice to make the dumb communicate with other people through voice . The Arabic words gesture as shown in Figure 1.

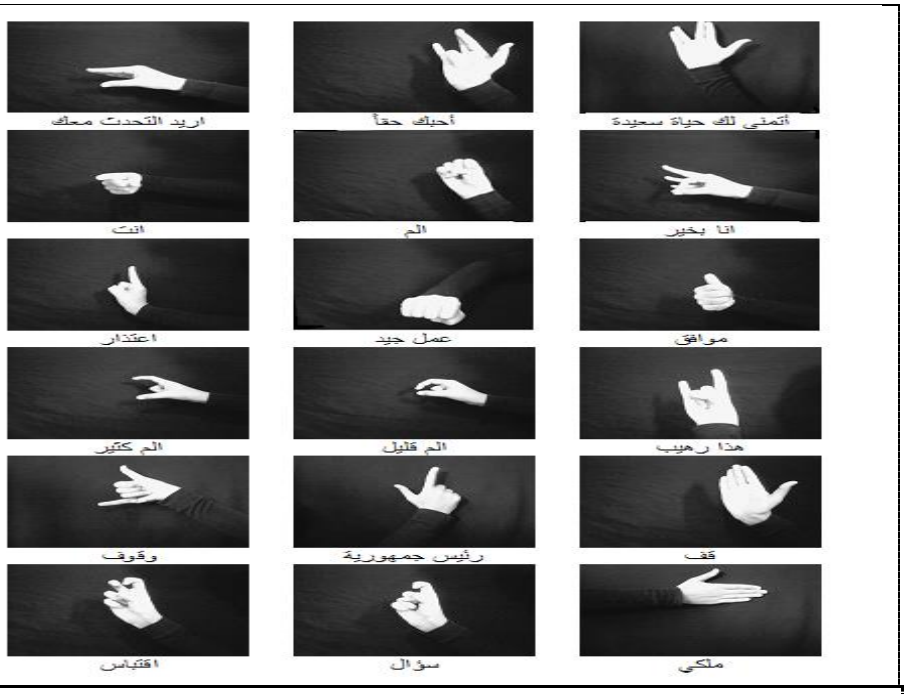


Figure 1. Gesture of Arabic Words

2.Clustering Algorithms and Classification Algorithms

There are many algorithms for clustering and classification, in our algorithm we tried to use the C4.5, K-mean, K- Medoid algorithms and ANN the result from our experiments shows that C4.5 is the best one and high accuracy.

C4.5 algorithm

C4.5 is a standard algorithm for inducing classification rules in the form of the decision tree. As an extension of ID3, the default criteria of choosing splitting attributes in C4.5 is information gain ratio instead of using information gain as that in ID3, information gain ratio avoids the bias of selecting attributes with many values[7].

C4.5. Algorithm steps [7]:

- Check the basic cases.
- For each calculate features :(Acquire the normalized information from the division on an attribute X).
- Select the best features that have the highest gain for information.
- Create a node is divided by the best decision point, such as the root node.
- Repeated the sub-menus obtained by splitting on the best a and adding those nodes as the children node.

3.Proposed Algorithm

The proposed algorithm consists of four steps (images acquisition, preprocessing step, features extraction, classify (in training) or comparison (in testing) and convert into voice). as show in Figure 2.

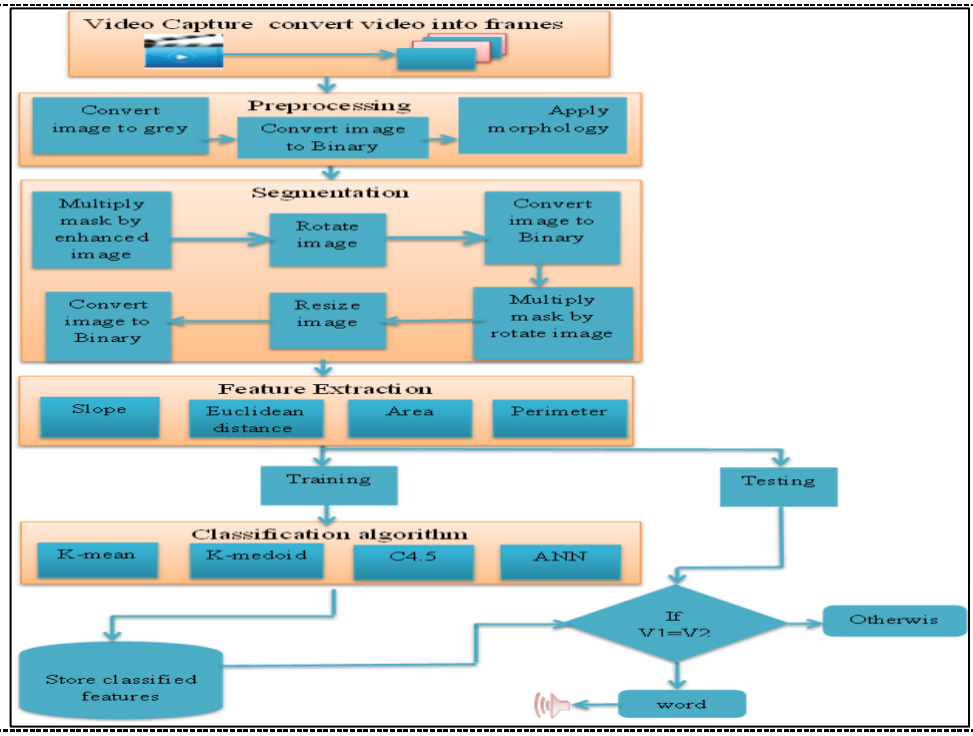


Figure 2. General Block Diagram of Proposed System

3.1 Dataset Aquisyion

We created our dataset by using an external camera in the laboratory by using different cameras to create our dataset. The background of images must be black color to be easy for classify objects (the image must contain only sign part). The position of the camera is also important issues to remove the background and keep the sing only.

Here we took eighteen words in Arabic (عمل, موافق, انت, الم,انا بخير, اريد التحدث معك, أحبك حقاً, أتمنى لك حياة سعيدة, اقتباس, سؤال,ملكي,وقوف,رئيس جمهورية,قف, الم كثير,الم قليل, هذا رهيب, اعتذار, جيد) words of three different persons, where these words will be as different templates. The videos are contains a series of frames (images) with size (720*1280 pixels) and (640 * 480 pixels). In this paper we used three videos contains 123 frames (images) with (720*1280 pixels) and 127 frames (images) with size (640 * 480 pixels) (for training stages).

3.2. Pre- Processingre

The Pre-processing includes the following steps:

- Transform the videos into the required frames (images $_k$)
- Convert the image (images k) to gray scale format then convert the resulted image into a double image the resulting image is an enhanced image.
- Transform the enhanced image (images $_k$) into a binary image.
- Remove little objects from the binary image using morphological operations (Dilation and closing).

3.3 Segmentation

- Segment the palm area from the resulting image of the morphological operations (Dilation and closing).
- Use the resulted image as a mask
- Multiply the enhanced image by mask.
- Calculate the angle according to the following Equation (1) [8].

$$\theta = \tan^{-1}((t_1 - t_2)/(1 + t_1 * t_2)) \qquad (1)$$

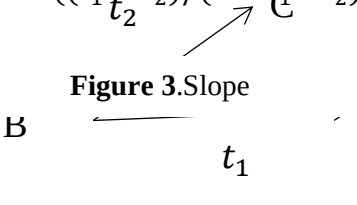


Figure 3.Slope

$t_1$is the slope between B and A, t_2 is the slope between B and C.

- Rotate (image $_k$) according to the following equation (2) [9].

$$\left.\begin{aligned}\hat{z} &= z * cos(\theta) - w * sin(\theta)\\ \hat{w} &= z * sin(\theta) - w * \cos(\theta)\end{aligned}\right\} \qquad (2)$$

(z, w) and $(\hat{z}, \hat{w})$: are pixel coordinates before and after rotation, respectively,
θ : is the counter clockwise angle of rotation.

- Convert the image (image $_k$) to a binary image and segment the palm area (used as mask).
- Multiply rotate (image $_k$)with a palm-sized mask.
- Resize the image [any * 100].
- Convert the image (Image $_K$) to a binary image.

3.4 Feature Extraction

In our proposed algorithm there are 18 geometrics features for each frame (image) these 18 features divided in (8-features) calculated from the distance between the 8 points within the center of hand, (8-features) calculated from the slop of the same 8 points in first (8-features), one feature calculated from area and one feature calculated from perimeter. We can calculate the 18-features by using as following steps:

- Extract 8 points of palm.
- Calculating the 8-features include the distance of 8 points by using Euclidean distance from the center of palm to the 8- points according to Equation (3) [10], and calculate the length of the palm and divide the distance by the length of the palm

$$DE_{zo} = \sqrt{\sum_{i=1}^{n}(x_{zi} - x_{\epsilon} \quad (3)}$$

Where as:
n: Number of properties
DE_{zo}: Distance between points and center of palm
x_{zi} : The coordinates of the i property for Z (where Z: points)
x_{oi}: The coordinates of the i property for o (where o: center point of palm)

- The second 8-features include the slop from center of palm to the 8-points as according to Equation (4) [8].Then calculate ($\tan^{-1}$) of the slope.

$$Slop = \frac{yz - yo}{xz - xo} \qquad (4)$$

Where as:
yz: the y- axis points
x_z: the x- axis points
y_o: the y- axis of center value point of palm
x_o: the x- axis of center value point of palm

- Calculate center point of palm as in Equation (5) [11].

$$x_o = \frac{\sum x_{oi} A_i}{\sum A_i}, \qquad y_o = \frac{\sum y_{oi} A_i}{\sum A_i} \qquad (5)$$

Where:
x_o: the x- axis of center value point of palm
y_o: the y- axis of center value point of palm
x_{oi}: The distance at which the center of the shape moves away from the junction point of the axes on the axis (x)
y_{oi}: The distance at which the center of the shape moves away from the junction point of the axes on the axis (y)
A_i: Area the shape

- Calculate area of palm as according to Equation (6) [12], and divided the results by 10000 (to reduce the big numbers).

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i \times y_{i+1}) - (x_{i+1} \times y_i) \qquad (6)$$

n: Number of points

x_i : x- axis coordinates points

y_i : y- axis coordinates points

- Calculate Perimeter of palm as according to Equation (7) [13] and divided the results by 500.

$$Per = \sum_{i=0}^{n-1} x_i \qquad (7)$$

n: Number of ribs

x: length of the rib

- Calculate the feature vector for each word (18 words) by using two steps:
- Calculate the average of features for the same words from different images of the same word.
- Calculate the feature vector for the first image of the word and ignore the rest images of the same word.

4. Result

The proposed algorithm contains two parts; one for training with 250 images while the second part is for testing by using 18 images as shown in Figure 2. In training part we need to calculating 18 features using C4.5 algorithm for classify the 18 types of words for each image as shown in Table 1 for distance, slop, area and perimeter respectively, then calculate the features for each words 18 types words as shown in Table 2 (the average of features for the same words from different images of the same word) and Table 3 shows the feature vector for the first image of the word and ignore the rest images of the same word. When we used three videos with 250 frames (images) we found that the results from four clustering algorithms; K-mean, K-mediod, C4.5, and ANN, for 18 words which gave different results for recognition as shown in Table 4 as following:

When we Implement of the k-mean cluster algorithm on the extracted features we found the accuracy of K-mean is 95.2000% , K-medoid is 91.9111% , C4.5 is 100% and ANN is 91.2727%for training stage when the dataset is 250, as shown in Table 4 and Figure 3 respectively.

Table 1 .Features vector of word

Feature Geometry																	
Distance								Slope								Area	Peri.
D1	D2	D3	D4	D5	D6	D7	D8	S1	S2	S3	S4	S5	S6	S7	S8		
0.1610	0.1808	0.1383	0.1376	0.2304	0.2325	0.2108	0.2205	-0.6508	-0.7832	-1.4638	-1.5361	-0.1267	-0.1836	1.1912	1.0927	0.4891	0.7902

Table 2 . Features vector of average 18 words

word	Number of image	Feature Geometry																	
		Distance								Slope								Area	Perimeter
		D1	D2	D3	D4	D5	D6	D7	D8	S1	S2	S3	S4	S5	S6	S7	S8		
وقف	24	0.2361	0.2339	0.1676	0.1595	0.2491	0.2548	0.2377	0.2542	0.3185	0.2368	-1.1806	-1.0451	-0.3602	-0.4138	0.9003	0.8236	0.5866	1.0113

ملكي	17	0.1652	0.1683	0.1930	0.1967	0.1359	0.1250	0.1524	0.1573	-0.1445	-0.2421	1.3325	1.2636	0.6734	0.2277	-0.5935	1.2345	0.4259	0.6585
قف	16	0.2506	0.2499	0.2213	0.2201	0.2248	0.2295	0.1393	0.1180	0.0753	-0.0249	-1.4027	-1.0462	-0.2291	-0.3086	-0.9949	1.0926	0.7258	0.8587

Table 3. Features vector for the first image of the word

word	Number of image	Feature Geometry																	
		Distance								Slope								Area	Perimeter
		D1	D2	D3	D4	D5	D6	D7	D8	S1	S2	S3	S4	S5	S6	S7	S8		
وقوف	24	0.2596	0.2485	0.1807	0.1739	0.2561	0.2643	0.2455	0.2594	0.7512	0.7022	-1.0083	-1.0732	-0.3891	-0.4592	0.8261	0.7694	0.5951	1.0999
ملكي	17	0.1704	0.1750	0.2005	0.2041	0.1399	0.1379	0.1492	0.1617	-0.2302	-0.3230	1.2279	1.1818	0.6577	0.6385	-1.3173	1.1052	0.4411	0.6786
قف	16	0.3339	0.3333	0.2114	0.2072	0.3048	0.3082	0.1678	0.1436	0.0608	-0.0091	-1.1139	-1.1573	-0.1768	-0.2301	-1.0274	-1.5702	1.2197	1.0646

Table 4 .shown the difference between four algorithms rate

word	No image	C4.5	K-mean	K- mediod	ANN
	250	100%	95.2000%	91.9111%	91.2727%

As shown in Table .4 the C4.5 algorithm is the best algorithm in the classification and accuracy

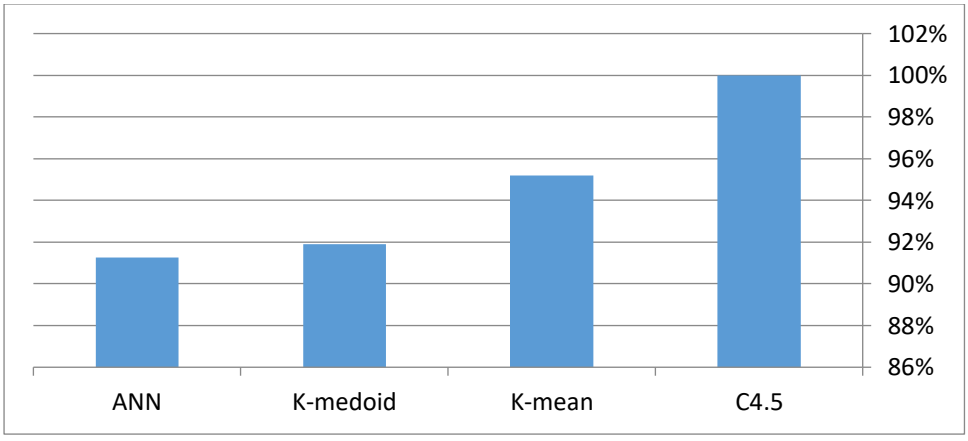


Figure 3. shown the difference between Four algorithms rate

In testing stage the features will classify where the input image will be in cluster or class of 18 types of words by comparing the feature vector with the 250 vectors stored in the dataset then the result will convert the class type or cluster type into corresponding voice (words) as shown in Figure 2. The testing of our algorithm is done by using Equation (8)[14] which is the modify of the Standardized Euclidean distance , by using Equation (3) of the Euclidean distance and also by using Equation (9) [15] of the correlation to compare the new features of input image with the classified features database of images.The results shows the accuracy of three ways (Modify of the Standardized Euclidean distance, Euclidean distance and Correlation) in Tables 5 and Figure 4

$$\mathrm{Dst} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sqrt{\frac{1}{(n-1)}\sum_{j=1}^{n}(x_j - \bar{x})^2 + (y_j - \bar{x})^2}}} \quad (8)$$

Where :

x_i: is the i^{th} valuse of first vector value.
y_i: is the i^{th} valuse of second vector value.
n: is the number of elements in vector.
$\bar{x}$: *is the mean value of first and second vector.*

$$R1 = \frac{\sum_i (x_i - x_m)(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2}\sqrt{\sum_i (y_i - y_m)^2}} \quad (9)$$

Where xi is the intensity of the ith value in vector 1, yi is the intensity of the ith value in vector 2, xm is the mean intensity of vector 1, and my is the mean intensity of vector2.

Table 5. the result accuracy of by using three ways

No. of tested images	No. of dataset images	Accuracy by using		
		Euclidean distance	correlation	Modify of the Standardized Euclidean distance
18	250	94.4444%	94.4444%	100%

The result shows that Modify of the Standardized Euclidean distance is the best accuracy from Euclidian distance and the correlation.

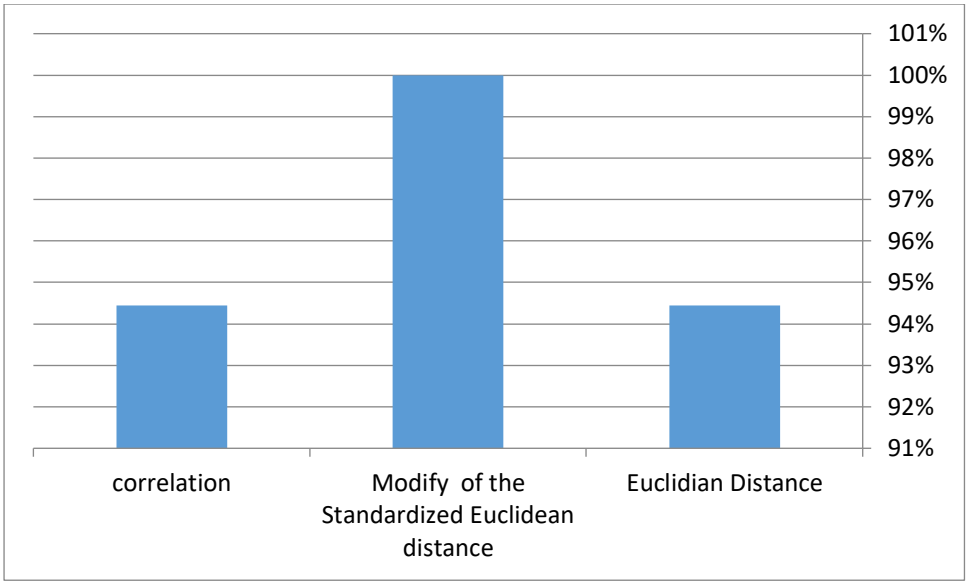


Figure 4.shows the accuracy rate using (Euclidian Distance ,Correlation and Modify of the Standardized Euclidean distance

5.Conclusion

In this paper, we have designed a system for recognition of Arabic words for gesture language based on clustering methods. In our experiments, we found that the geometric features (distance, area, perimeter, and slope) are the good eatures rather than the others features such as (shape, texture, color,…..etc). There are many clustering and classify algorithms used in our proposed algorithm such as (K-mean, K-medoid, C4.5, and ANN) where the experiments found that C4.5 algorithm is the best one in clustering or classify with the percentage of(100%) in training and (100%) in testing. In the testing stage, the results found that the modify of the Standardized Euclidean distance is best metric for calculating the corresponding features vector of the tested image to know the type of which words (18 words) while others metrics such as (Euclidean distance and Correlation) is lower accuracy than the Modify of the Standardized Euclidean distance.

References :

[1] Ch. V.N. Syam Babu, V.J.K. Kishor Sonti and Y. Varthamanan 2016 *Design and Simulation of Communication Aid for Disabled Using Threshold Based Segmentation* (I J C T A Vol 9,No7) pp. 3275-3281.

[2] Mohamed S. Abdalla and Elsayed E. Hemayed 2013 *Dynamic Hand Gesture Recognition of Arabic Sign Language using Hand Motion Trajectory Features*(Global Journal of Computer Science and Technology Graphics & Vision Vol 13, Issue 5)pp26-33.

[3] Hemina Bhavsar and Dr. Jeegar Trivedi 2017 *Review on Classification Methods used in Image based Sign Language Recognition System*(International Journal on Recent and Innovation Trends in Computing and Communication Vol5 ,Issue 5)pp 949 – 959.

[4] Kumud Tripathi ,Neha Baranwal and G. C. Nandi *2015Continuous Indian Sign Language Gesture Recognition and Sentence Formation*(Procedia Computer Science Vol 54)pp 523 – 531.

[5] Liu Yun, Zhang Lifeng and Zhang Shujun 2012 *A Hand Gesture Recognition Method Based on Multi-Feature Fusion and Template Matching* (Procedia Engineering Vol 29)pp 1678 – 1684.

[6] Pablo Barros, Nestor T. Maciel-Junior , BrunoJ.T. Fernandes , ByronL.D. Bezerra and Sergio M.M. Fernandes *2017A dynamic gesture recognition and prediction system using the convexity approach*(Computer Vision and Image Understanding Vol 155)pp139-149,2017.

[7] Wei Dai and Wei Ji 2014 *A Map Reduce Implementation of C4.5 Decision Tree Algorithm*(international journal of database theory and application Vol 7,No 1)pp.49-60..

[8] Christopher Clapham and James Nicholson 2009 *Oxford Concise Dictionary of Mathematics*(OUP

oxford).
[9] Hermann K. 2011 *Real-Time Systems Design Principles for Distributed Embedded Applications*(Springer) Second edition.
[10] Michel Marie Deza and Elena Deza 2009 *Encyclopedia of Distances*(Springer)pp 94.
[11] Dan B. Marghitu and Mihai Dupac 2012 *Advanced Dynamics*(Springer)Chapter 2 pp 73-141.
[12] H. Stroud 1971 *Approximate calculation of multiple integrals*(Prentice-Hall Inc., Englewood Cli_s, N. J.).
[13] Dr Yeap Ban Har,Dr Joseph Yeo,Teh Keng Seng,Loh Cheng Yee,Ivy Chow,Neo Chai Meng and Jacinth Liew 2018 *NEW SYLLABUS MATHEMATICS TEACHER'S RESOURCE BOOK1*(OXFORD UNIVERSITY Press) 7th edition.
[14] Pavol OR.ANSKÝ 2009 *Fundamentals of Mathematical Statistics*(Slovakia: Statistics Faculty of Management, University of Presov)
[15] A. Miranda Neto, A. Correa Victorino, I. Fantoni, D. E. Zampieri, J. V. Ferreira and D. A. Lima 2013 *Image Processing Using Pearson's Correlation Coefficient: Applications on Autonomous Robotics*(IEEE).

Image Classification based on CBIR

Shaker K. Ali[1] **Salah A. Sadoon[2]**

1Computer sciences and mathematics college, University of Thi_Qar, Thi_Qar, Iraq.
2College of Education for Pure, University of Thi_Qar, Thi_Qar, Iraq.
shaker@utq.edu.iq and abdsalah847@gmail.com

Abstract.. In this paper, we present a new way to classify four types of images (Car accidents, Fire, Abnormal objects in street and Digs) which will be sent to four government places; Civil Defence, police station and Municipal. The classification method depends on the Content-Based Image Retrieval (CBIR), where we use a new method. In this method, we use a combination of three methods to extract features from an image; Single Value Decomposition (SVD), Edge Histogram Descriptor (EHD) and Color Auto-Correlogram for Extraction Features. You will use these features to find the closest similarities to the query image from the database images by selecting the closest 3 images, then choosing the class to which the closest two images belong to the retrieved. The combined method showed 100% accuracy in training phase and 100% test phase accuracy.
Keywords: Classification, CBIR, SVD, EHD and Color Auto-correlogram.

1.Introduction

In general, classification is the process of sorting things, i.e. reclassifying those things into classes for each class with similar characteristics. Image classification is the identification of the classes to which the image belongs from one of the predefined classes. Image classification requires several accurate calculations and the use of suitable tools, depending on the type of application. Accuracy of classification comes from the use of an appropriate method and tool. We will use the CBIR system to classification the image. CBIR system is to restore the image according to the visual content of the images in the database, in other words according to the shape, color and texture[1].The main task of CBIR systems is to extract image features and features to compare similarities with another image and to define a comparison rule, as it depends on the pixel values in the image. When measuring similarity between images, the images are represented by these features. [2]. Feature extraction is the vector formation stage based on image data, in other words converting the original image data into measurable data. Images have rudimentary features such as shape, color and texture. In CBIR, the image is converted to a feature vector and compared to the images in the database to retrieve the most similar images to the query image. [3].Content-based image retrieval uses mathematical metrics to measure similarities, to show similarities between the two images, where distance laws such as Euclidean distance and Manhattan can be used. To complete the comparison process, this requires obtaining image features, as several methods can be used to extract features such as Singular value decomposition (SVD) and The edge histogram descriptor (EHD) to obtain the shape feature, using color moment to obtain the color feature.

2. Image Features:

The stage of obtaining the image features is an important stage in image retrieval. The system's performance depends mainly on the efficiency of these extracted features to describe the Image components. The feature is data of numerical value extracted from image, that are difficult to understand by humans. The features extracted from any image are less than the original image data and with a bigger difference. This reduces the overhead of handling the image set. The image contains two types of features is global and local features. General feature is used to describe the image as a whole, as they are used to retrieve images, reveal objects, and categorize. But local feature is limited to describing image corrections that are used to identify and identify objects. Using global and local features together helps increase recognition accuracy, but with an increase in computational costs [4]. Any image contains many features

that can be extracted and relied upon to describe that image, from those features are color, shape and texture that have been studied and used effectively and widely in CBIR systems [5].

2.1. Singular Value Decomposition

An image is a set of numbers stored in rows and columns as an matrix. The numbers represent the image information, which it keeps as a table. Linear algebra is the mathematical study of matrices.

Matrix analysis and processing helps to analyze large pieces of data in an easier way, with each pixel of the image being represented as a number in the matrix. The columns and rows in an matrix contain the position of that value relative to the position of the image [6]. SVD is used to minimize the large dimensions of original data to a lower dimensional space where the substructure of the original data is shown more clearly and orders it from most variation to the least. By using SVD the area of most variation can be found and its dimensions can be minimized. In other words, SVD is considered as a method for data minimizing [7]. Single-value Decomposition (SVD) is a method that deals with the matrices at linear algebra. It decompose the matrix into three matrices of different dimensions that contain the original matrix information. two matrices, the left singular vectors and the right singular vectors consisting of single vectors give information about the structure of the original matrix and an matrix consist of singular values describing the strength of the specific components of the original matrix. [8]. SVD according to the theory of linear algebra states: a Matrix m×n rectangular A can have m rows and n columns in them dissociation of three matrices, as given in Equation 1 [9]

$$A = USV^T \quad (1)$$

Where, U is the left singular vectors matrix of A, and it a m×n matrix of the orthonormal eigenvectors of $\boldsymbol{AA^T}$, $\boldsymbol{V^T}$ is the right singular vectors matrix of A, it the transpose of a n×n matrix containing the orthonormal eigenvectors of A , and are the identity matrices of size n and p, which achieve the two Equations 2 and 3 respectively, and S is the singular values matrix of A, and it a n×n diagonal matrix contains nonnegative singular values are the square roots of the eigenvalues of A, which given in Equation 4 [9].

$$U^T U = I_{n\times n} \quad (2)$$

$$V^T V = I_{p\times p} \quad (3)$$

$$S = \begin{bmatrix} \sigma_1 & 0 & \cdot & 0 & 0 \\ 0 & \sigma_2 & \cdot & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \sigma_{n-1} & 0 \\ 0 & 0 & \cdot & 0 & \sigma_n \end{bmatrix} \quad (4)$$

Where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$, $p = \min\{m, n\}$, and $U = [u_1 \quad \dots \quad u_m], V = [v_1 \quad \dots \quad v_n]$.

To perform singular value decomposition, the eigenvectors and eigenvalues of matrices must be obtained $\boldsymbol{A^T A}$ and $\boldsymbol{AA^T}$. The columns of V are the eigenvectors of $\boldsymbol{A^T A}$.such that, the matrix $\boldsymbol{A^T A}$ it is as follows:

$$A^T A = USV^T VSU^T = US^2U^T \quad (5)$$

The columns of U are the eigenvectors of $\boldsymbol{AA^T}$. The matrix $\boldsymbol{AA^T}$ it is as follows:

$$AA^T = USV^T VSU^T = US^2U^T \quad (6)$$

The last matrix is S which contains the singular values of the matrix $\boldsymbol{A^T A}$ or $\boldsymbol{AA^T}$ eigenvalues. The base diameter of the S matrix contains singular values in descending order, it are real numbers. If A is a matrix with real values then the values in V and U are also real.In our proposed method we used 7 left Single vectors(r=7) and 3 right singular vectors (c=3).Because this gives the best results, The sum of the total values of SVD is 21 (features) (r×c=21)

2.2 . Edge Histogram Descriptor (EHD)

The shape is one of image features an important role in identifying things in the image, and shape features help to identify the object within the image[10]. Edge histogram descriptor (EHD) is one of The methods widely used to detect shapes in images, which describes the relative frequency of the 5 types of edges in the image [10]. The histogram is the most used feature to describe the global features of an image. It is more powerful constancy for translation, image rotation and normalization, It leads to the expansion of stability. These characteristics make it more useful in indexing and retrieval of images. The histogram is the main tool used in EHD technology to determine the types of edges in an image [10,11]. EHD is a process that describes 5 types of edges in each local area of the image called a sub-image and how those edges are distributed. See figure 1, this is done by divide the area of the image into 4x4 of non-overlapping blocks, one block is a sub-image of the original image. The result of the splitting process is 16 sub-images of equal size, without taking into account the original image size [10,11].

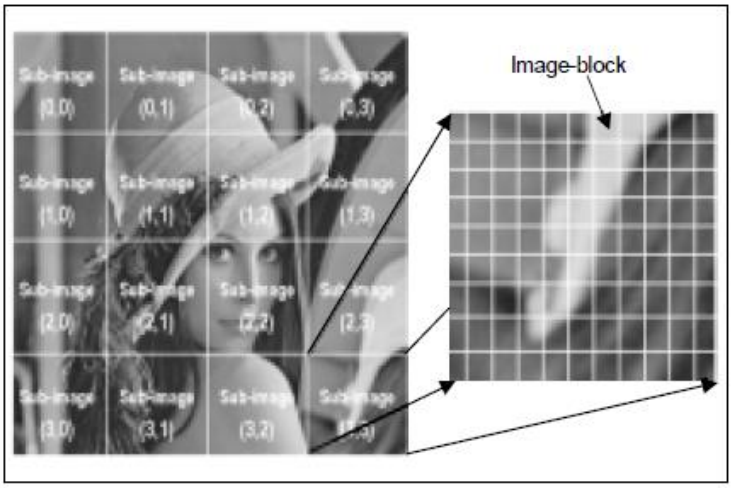


Figure 1. Image-block (sub-image).

The histogram is generated each edge in the sub image. Each sub image contains 5 types of edges: are vertical edge, horizontal edge, 45 degree diagonal edge, 135 degree diagonal edge, and non-directional edge .Figure 2. The sub-image histogram is the relative frequency of occurrence of these types of edges in the sub-image.

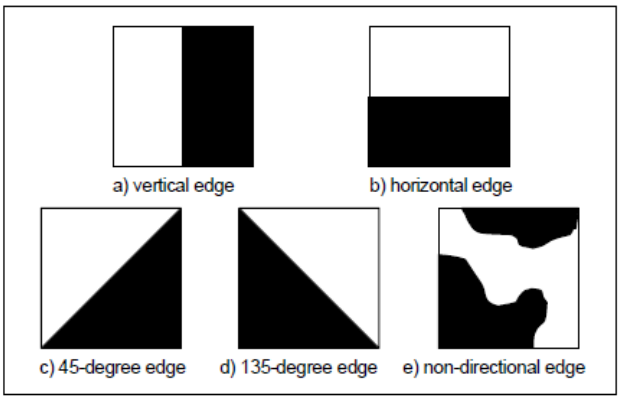


Figure 2. Five types of edges in a sub- images

Each local histogram contains 5 bins, as shown in Figure 3. Each bin corresponds to one of the five types of edges. According to the division of the image into 16 image block (sub-images) and each image block has 5 bins of histogram, the result is a total of 80 histogram bins.The 80 histogram bins has its own indications in terms of location and edge type as in Figure 4.

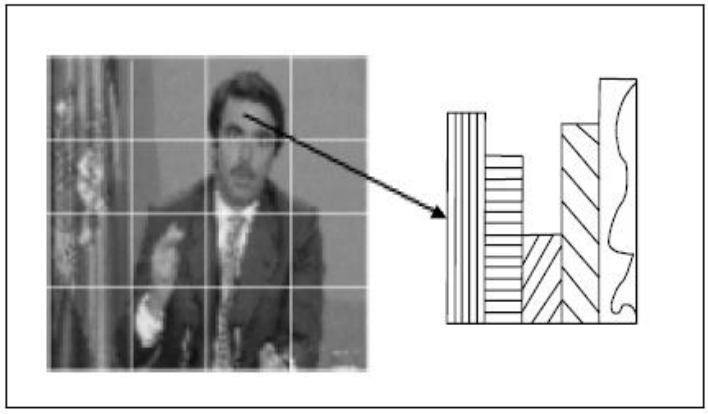


Figure 3. Five types of edge in sub-image

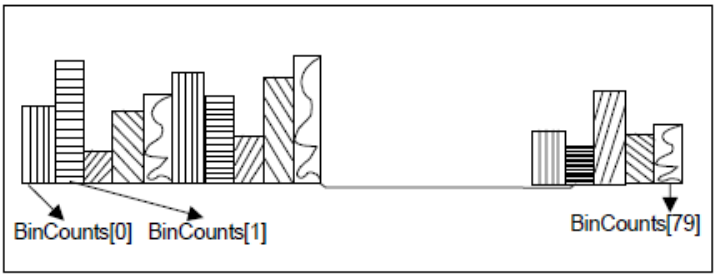


Figure 4. 1x80 matrix of 80 bins

The sub-image is visited from (0.0) to the sub-image (3,3), the bitmap and local graph are done for all 16 sub-pictures, and on this basis the arrangement is made bins of each sub-image Figure 5. The types of edges are arranged as follows: Vertical edge, horizontal edge , 45 degree diagonal edge, 135 degree diagonal edge and non-directional edge [10]. See also Table 1 summarizing semantics of 80 bin [11,12].

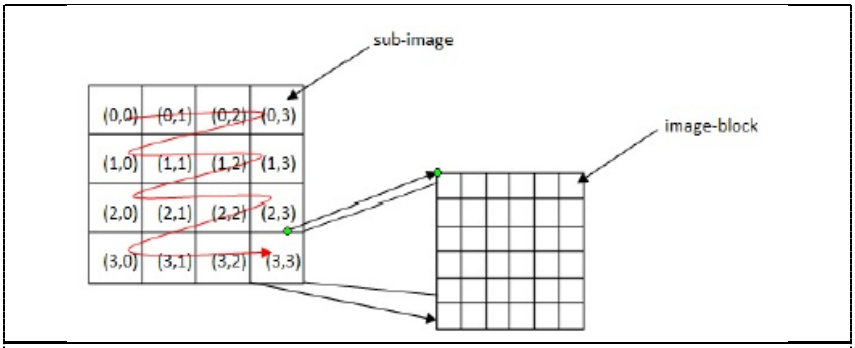


Figure 5. Image visit and definition of sub picture in EHD

Table 1. Semantics of local edge bins.

Histogram bins	Semantics
BinCounts[0]	Vertical edge of sub-image at (0,0)
BinCounts[1]	Horizontal edge of sub-image at (0,0)
BinCounts[2]	45-degree edge of sub-image at (0,0)
BinCounts[3]	135-degree edge of sub-image at (0,0)
BinCounts[4]	Non-directional edge of sub-image at (0,0)
BinCounts[5]	Vertical edge of sub-image at (0,1)
:	:
BinCounts[74]	Non-directional edge of sub-image at (3,2)
BinCounts[75]	Vertical edge of sub-image at (3,3)
BinCounts[76]	Horizontal edge of sub-image at (3,3)
BinCounts[77]	45-degree edge of sub-image at (3,3)
BinCounts[78]	135-degree edge of sub-image at (3,3)
BinCounts[79]	Non-directional edge of sub-image at (3,3)

Table 2 shows the full indications for EHD with 80 histogram boxes. Image block is a basic unit for extracting edge information, and each box is divided by the total number of image blocks in the sub image to be set. Through the mass of the image it becomes clear to us whether there is an edge or not and what is the prevailing edge of the 5 types of edges. Then find the value of the edge histogram, and increase the edge of the opposite edge by one. Otherwise considered monotonous area in the image (i.e. when there is no edge). This particular block of images does not contribute to any of the 5 edge boxes.In the same way, we can find another 70 bins of the same image, by finding 5 types of edges for the whole image (global). And find 65 bins by grouping the image-blocks as in Figure 6 by 13 groups and then finding 5 types of edges for each group (semi-global). As a result, the total number of bins is 150, through which the image can be matched with another image.

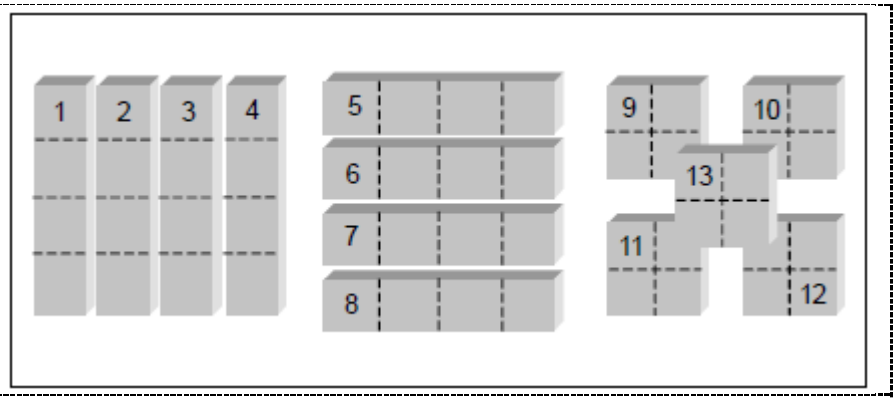


Figure 6. Grouping of sub-images for semi-global histograms

2. 3. Color Auto Correlogram

It is the image feature that can be extracted to denote (as a description) the image color feature, which shows the spatial correlation of color at a distance.A correlogram is an expression of the correlative distribution of color in an image. It describes the spatial correlation of a pair of colors changing with distance. As shown in Figure 7 [13].

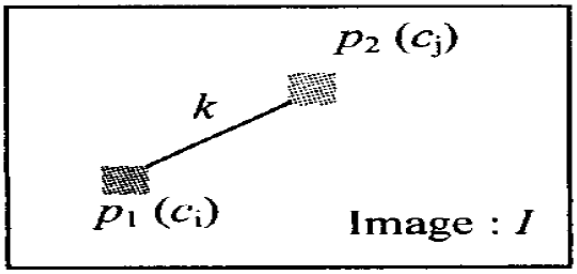


Figure 7. Spatial correlation of a pair of colors in a distance.

Correlogram is storage table indexed by pairs of Colours (ci, cj) where d-th entry is likely to be found pixel color cj from pixel color ci in distance d. Whereas, the color auto correlogram is a subset of the correlogram, which is a table indexed by one color where the entry is d-th The probability of finding a ci color pixel from the same pixel In the distance d. Thus it demonstrates spatial automatic correlation The relationship between identical colors only [14,15].

Let I be an n x n image. And ci, cj ϵ {1,2,.....,m}. Suppose the distance d ϵ {1,2,.....,n}. The correlogram $\boldsymbol{\gamma}_{\mathbf{ci,cj}}^{(\mathbf{d})}$ of image I for color pair (ci, cj) ϵ {1,2,.....,m} and d ϵ {1,2,.....m} is defined as :

$$\boldsymbol{\gamma}_{ci,cj}^{(d)}(I) = \mathbf{Pr}_{p1\in I_{ci},p2\in I}\left[\boldsymbol{p2} \in I_{cj} \middle| |\boldsymbol{p1}-\boldsymbol{p2}| = d\right] \qquad (7)$$

The auto-correlogram with color ci and a distance d, it is defined of image I as follow:

$$\alpha_{ci}^{(d)}(I) = \gamma_{ci,ci}^{(d)}(I) \qquad (8)$$

In another Equation :

$$\alpha_{ci}^{(d)}(I) = \Pr[|p1-p2| = d, p2 \in I_{ci} | p1 \in I_{ci}] \qquad (9)$$

Auto correlogram collects space information and color information for each pixel in the image, which requires visiting all neighbours of this pixel. It shows only the matching colors in terms of spatial correlation in the image [13] .

3. proposed system

The goal of the proposed system is to design an effective system for categorizing 4 types of images. This system relies primarily on CBIR to retrieve the image. CBIR also is a proposed new recovery system based on the integration of the SVD, EHD and Color Auto -gorrelogram features described above. Our proposed system showed satisfactory results using the following algorithm. As show flowing Figure 8.

Algorithm for image classification based on CBIR using Multi Features.

Step 1: Load the image database in the Mat lab workspace.

Step 2: For each image from the database, resize the image size to 116 x 116.

Step 3: For each image from the database, apply mean filter 3x3.

Step 4: For each image from the database, extract SVD, EHD and Color Auto- Correlogram features .

Step 5 : Store the features database obtained from previous steps. Each image is represented by a vector of 235 features.

Step 6 : Load the query image.

Step 7: Repeat steps 2-4 on the query image.

Step 8: Measuring the Euclidean distance between the feature vector of the query image and all of the feature database vectors.

Step 9: Retrieve 3 images that generate the lowest distance from the query image.

Step 10 : The query image is from the class to which the two nearest images belong.

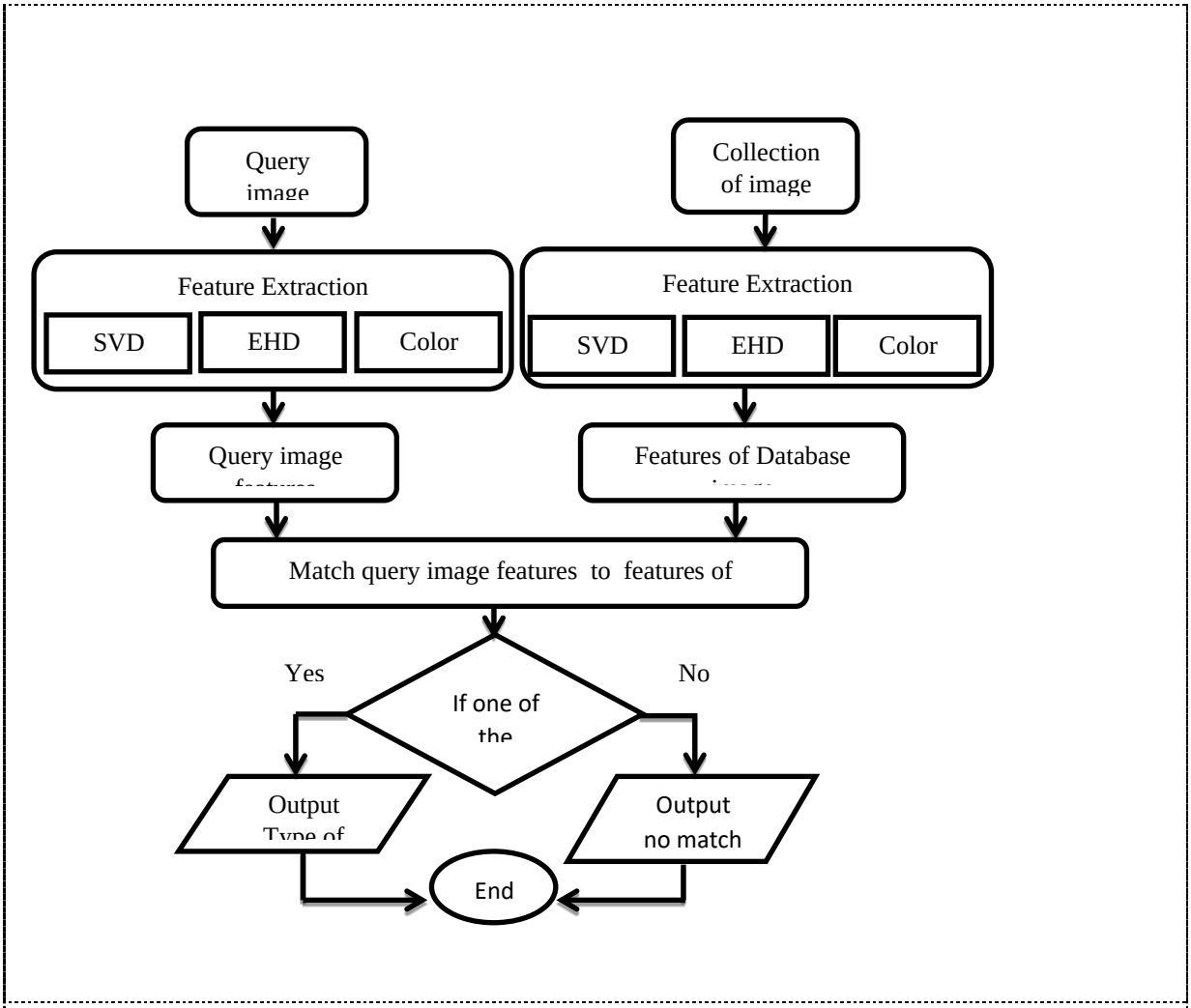


Figure 8. proposed system.

The Euclidean distance to calculate the minimum distance between two vectors such as (q, p) is calculated as follows:

$$D(q,p) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (10)$$

4.Experiment and results:

In the experiment, we used a training database consisting of 4 classes of images. The total number of database images is 400 images. 100 images for each class.

The test images are 100 images 25 of each class. All images were collected from the internet and phone database. The process of classifying any training image or test image is done on the basis of retrieving the closest 3 images from the database to the query image. When retrieving two images from the same class, the query images belong to that class. We conducted the experiment twice, on the first attempt, the classification was based on retrieving the closest image to the query image. Using the SVD, EHD, and Color Auto-correlogram features separately, then combine these features, as the result is as shown in Table (2).

In the second attempt, the classification was done on the basis of retrieving the three closest images of the query image, using the SVD, EHD, and Color Auto-correlogram features separately, then combine these features . The result was as shown in Table (3).

Table 2. Retrieval closest 1 images.

Classes	SVD	EHD	Color Auto-correlogram	SVD+EHD+Color Auto-correlogram
Class 1	84%	76%	40%	84%
Class 2	88%	56%	28%	96%
Class 3	92%	68%	12%	92%
Class 4	88%	60%	44%	92%
Avg.	88%	65%	31%	91%

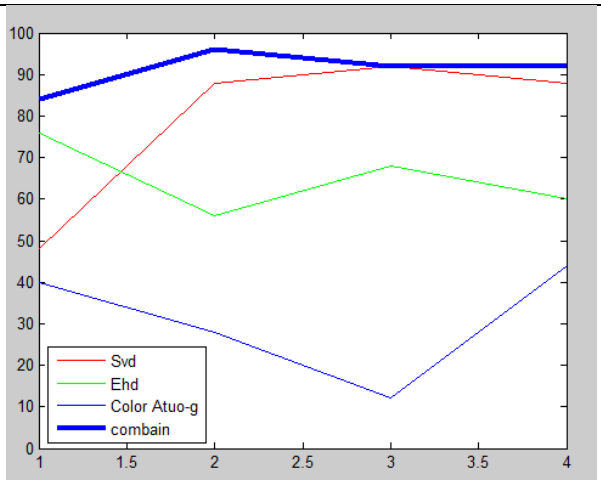


Figure 9. Comparison of SVD, EHD, and Auto-Correlgram and combine if the closest image is retrieved

Table 3.Retrieval closest 3 images.

Classes	SVD	EHD	Color Auto-correlogram	SVD+EHD+Color Auto-correlogram
Class 1	92%	84%	8%	100%
Class 2	92%	44%	16%	100%
Class 3	92%	76%	8%	100%
Class 4	88%	52%	36%	100%
Avg.	91%	64%	32.25%	100%

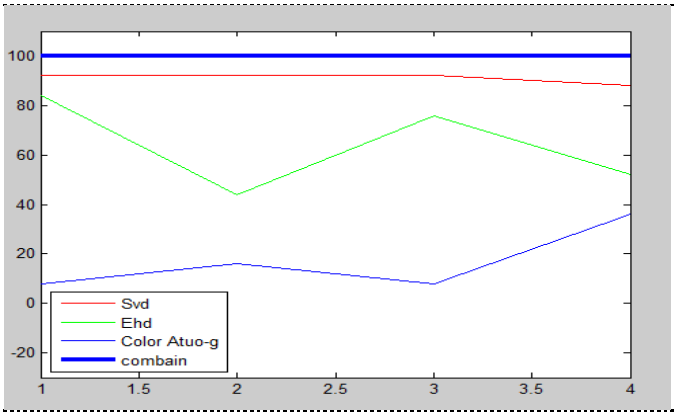


Figure 10. Comparison of SVD, EHD, and Auto-Correlgram and combine if the closest three image is retrieved

The results above show improvement of the classification process when using the option to retrieve to the nearest 3 images, as shown by the strength of the use of SVD and EHD

5 . Conclusion

In this paper, we have proposed a method for image classification based on CBIR system. The proposed CBIR system is based on three descriptors: SVD, EHD, and Color Auto-Correlogram. 21 features are extracted using SVD from the image, 150 using EHD, and 64 using Auto-Correlogram. The total is 235 features for each image in the database. They are matched with the image of the query using Euclidean Distance.

Reference

[1] Reshma Chaudhari and A. M. Patil 2012 Content Based Image Retrieval Using Color and Shape Features(International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering ,Vol.1, Issue 5)

[2] Gunjan Khosla, Dr. Navin Rajpal and Jasvinder Singh 2014 Evaluation of Euclidean and Manhanttan Metrics In Content Based Image Retrieval System(Journal of Engineering Research and Applications ,Vol 4, Issue 9) pp.43-49.

[3] Vadhri Suryanarayana, Dr. M.V.L.N. Raja Rao, Dr. P. Bhaskara Reddy and Dr. G. Ravindra Babu 2012 IMAGE RETRIEVAL SYSTEM USING HYBRID FEATURE EXTRACTION TECHNIQUE(International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 1).

[4] Abdulkadhem Abdulkareem Abdulkadhem, 2019 An Intelligent Roads Map Discovery based on Video Tracking Techniques(University of Babylon, College of Information Technology).

[5] David G. Lowe 2004 Distinctive Image Features from Scale-Invariant Keypoints(International Journal of Computer Vision)

[6] Miss Samruddhi Kahu and Ms. Reena Rahate 2013 Image Compression using Singular Value Decomposition (International Journal of Advancements in Research & Technology, Volume 2, Issue 8).

[7] K. Mounika, D. Sri Navya Lakshmi, K. Alekya 2015 SVD BASED IMAGE COMPRESSION(International Journal of Engineering Research and General Science Volume 3, Issue 2).

[8] Sura Ramzi Sheriff 2010 Digital ImageWatermarking Using Singular Value Decomposition(Raf. J. of Comp. & Math's. Vol 7, No 3)

[9] Assad Hussein Thary 2014 Satellite Image Classification Using K-Means and SVD Techniques(B.Sc. in Computer Science / College of Science / Al-Nahrain University)

[10] Neetesh Prajapati, Amit Kumar Nandanwar and G.S. Prajapati 2016 Edge Histogram Descriptor, Geometric Moment and Sobel Edge Detector Combined Features Based Object Recognition and Retrieval System(International Journal of Computer Science and Information Technologies Vol 7 ,No 1) pp 407-412.

[11] Chee Sun Won, Dong Kwon Park, and Soo-Jun Park 2002 Efficient Use of MPEG-7 Edge Histogram Descriptor(ETRI Journal, Volume 24, No 1).

[12] Minyoung Eom, and Yoonsik Choe 2007 Fast Extraction of Edge Histogram in DCT Domain based on MPEG7(World Academy of Science, Engineering and Technology Vol 9).

[13] C. Umamaheswari, Dr. R. Bhavani and 3Dr. K. Thirunadana Sikamani 2018 Texture and Color Feature Extraction from Ceramic Tiles for Various Flaws Detection Classification (International Journal on Future Revolution in Computer Science & Communication Engineering Vol 4)

[14] Dipankar Hazra 2011 Retrieval of Color Image using Color Correlogram and Wavelet Filters (Proc. of Int. Conf. on Advances in Computer Engineering)

[15] Shiv Raj Singh, Dr. Shruti Kohli 2015 Enhanced CBIR using Color Moments, HSV Histogram, Color Auto Correlogram, and Gabor Texture (International Journal of Computer Systems (ISSN: 2394-1065), Vol 2, Issue 5)

Hybrid K-means Clustering (HK): Cluster Assessment via Rand index

Zahraa Radhi Waad 1, Bahaa Hussein Taher 2

1college of Computer Science and Mathematics, University of Thi-Qar, Iraq.
2college of Computer Science and Information, Technology, University of Sumer, Iraq.
1hhaass441@gmail.com
[2]Ghrabiuk@gmail.com

Abstract. This paper introduces a hybrid K-means clustering method, this method has better results than the standard K-means method in terms of accuracy. In order to evaluate the hybrid algorithm, it is compared with the standard algorithm in terms of accuracy on **synthetic data with normal distribution** and real data sets for single hierarchical and average hierarchical with Euclidean and Manhattan distances. In this paper, we determine the number of clusters by using a ratio from 0.1 to 0.9 from the total number of original data. And also, we used the external (Rand index) criteria with the purposes to evaluate the results obtained from hybrid K- means clustering and standard K-means clustering.
Keywords: Hierarchical clustering, K- means clustering, Hybrid clustering, External validation.

Introduction

Clustering methods are classified into two main groups: hierarchical and partition clustering. The partition clustering has many algorithms, the K-means method is the most commonly used clustering method among others methods because it is a simple algorithm and it can be implemented quickly. However, the main disadvantage of the K- means method is determining k (number of clusters) prior to clustering and also it chooses randomly the initial points to solve these problems so we combine agglomerative hierarchical clustering method and K- means clustering method introduces a hybrid K-means clustering method. In hybrid hierarchical- K-means method (HK) the initial centroids and k (number of cluster) create from agglomerative hierarchical clustering method then the clustering was improved with K- means by K- means method [1],[2]. And the merging of agglomerative clustering method and K- means clustering method so that it creates better clusters and passes the cluster information from one to other. generally, a combination of two methods will improve clustering's accuracy [2], [3].

Related work:

In 2005, Bernard Chen et al split training data into two sections. The hierarchical method runs over one section of the dataset to get information of the data and K- means method runs over other section [1]. He Ying et al. offered the HK algorithm depend on PCA, this algorithm is to work on the total dataset (instead of two sections). In the first phase, it reduces the dimension of the dataset by uses PCA technology and then locate the initial center of cluster by implementing an agglomerative method. Finally, we get the results by implementing the k-means method [2]. In 2010, Li Zhang et al offered merged divisive and the agglomerative method to treat irreversibility of HK. Divisive method obtains many clusters by implementing K-means method at each layer and then uses the agglomerative method to merge clusters [4]. In 2016, Wenhua Liu et al. offered a new method called iHK. In fact, the iHK is an improved HK algorithm since it reduces the time complexity for HK by Normalization of data [5].

Distance Measurements:

Many clustering methods are based on measuring similarity and dissimilarity between data by calculating the distance between two data points [6].

Euclidean distance:

It represents the length of the straight-line connection to two points X and Y. If $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ are two points in n dimension, it is computed as:

$$d(X-Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

The Euclidean distance is sensitive to outliers [6].

Manhattan distance **(Cityblook):**
It represents the distance measured along directions that are parallel to the x and y-axes. It between two n-dimensional vectors $X=(x_{1,}\,x_2,\ldots,x_n)$ and $Y=\left(y_{1,}\,y_2,\ldots,y_n\right)$ is

$$d(X,Y)=\sum_{i=1}^{n}|x_i-y_i|$$

Where $|x_i-y_i|$ appears the absolute value of the difference between x_i and y_i.

Clustering Methods:
Agglomerative Hierarchical Methods (bottom-up):
The Hierarchical clustering provides clusters of high - quality and easy understanding of the data, however, it is costlier. we focus on the agglomerative (bottom-up) method in this paper. This method Starts with one cluster for each a data point and recursively combining. The combining continues until given one cluster [7].
The agglomerative algorithm consists of the following steps:

1.Start.
2. calculate the distance matrix between the data points.
3. let cluster for each a data point
4. integrate the two closest clusters and then change the distance matrix.
5. if more than one cluster remains then go to step 4.
6. End

We used two approaches of agglomerative clustering in this paper, Average and Single Linkage.

Single linkage:
It is one of the several methods of agglomerative hierarchical clustering. The distance between every pair of clusters is measured and two clusters ($A\ and\ B$) having the minimum distance are combined at each step in this method. The minimum distance between an object point in A and an object point in B represents The distance between cluster A and cluster B**.**
$D(A,B)\ =\ min\{d(y_i,y_j),for\ y_i\ in\ A\ and\ y_j\ in\ B\}$,where $d(y_i,y_j)$ appears the euclidean or another distance between the vectors $y_i and\ y_j$.

Average Linkage Clustering:
In average linkage, we define the distance between two clusters ($A\ and\ B$) $\mathrm{D(A,B)}=(\frac{1}{\mathrm{n_A\,n_B}})\sum_{\mathrm{i=1}}^{\mathrm{n_A}}\sum_{\mathrm{j=1}}^{\mathrm{n_B}}\mathrm{d(y_i,y_j)}$ where $n_A$and $\mathrm{n_B}$ are the number of points in A and B. If the clusters $A\ and\ B$ are merged using the centroid method, and if cluster A has more objects than B, then the new centroid.
$\bar{\mathrm{y}}_{\mathrm{AB}}\ =(\mathrm{n_A}\bar{\mathrm{y}}_{\mathrm{A}}\ +\mathrm{n_B}\ \bar{\mathrm{y}}_{\mathrm{B}}\ )/(\mathrm{n_A}\ +\mathrm{n_B}\ )$. May be much closer to $\bar{\mathrm{y}}_{\mathrm{A}}$ than to $\bar{\mathrm{y}}_{\mathrm{B}}$
Average linkage method can treat categorical and numeric data. But it can fail easily when clustering in hyper spherical shape and it is insensitive to outliers [7].

Partitioning Method(K-Means):

It is one of the most popular clustering methods and the best way to generating globular clusters. The K-means clustering is statistical, non-supervised, and iterative. It is used most widely in many areas like image segmentation, object recognition, etc. The objects can be moved between a cluster and another in this method but in the hierarchical methods that is not available. The aim of the K-means clustering finds the means of the clusters such that the distance between the data point to the cluster is minimized. So, the problem that we want to solve can be stated as [8]:

Input: k (is represent number cluster) and object x_1,x_2,....x_m

Minimize distortion $=\sum_{j=1}^{K}\sum_{i=1}^{m}\left|x_i^{(j)}-c_j\right|$ where $\left|x_i^{(j)}-c_j\right|$ It is represent distance measure between object $x_i^{(j)}$and center c_j .

The time complexity of k-means clustering is $O(n*k*d*i)$. When n represents the number of observations, k: number of clusters, i: number of iterations, and d: number of attributes [9],[10].

The algorithm consists of the following steps:

1. Start
2. Input: n objects.
3. Randomly k initial group centroids.
4. Repeat
5. Assign each object to the closest centroid.
6. Recalculate the positions of the k centroids.
7. Until Stopping Criteria.
8. End.

Stopping Criteria of K-means method:

- When the squared error is less than some small threshold value.
- No change in the members of all clusters.

Hybrid K-means Clustering Algorithm (HK):

Combine agglomerative hierarchical clustering method and K-means clustering method introduces a hybrid K-means clustering algorithm, which is the algorithm that has better results than the standard k-means algorithm in terms of accuracy. In hybrid K-means algorithm (HK) the k (number of clusters) and initial centroids create from agglomerative hierarchical method then the clustering was improved with K-means by K-means method [1], [2].

This hybrid hierarchical Clustering Method is implemented as follows:

- Do hierarchical clustering and cut the tree into k- clusters. The k- clusters equal to ratios from 0.1 to 0.9 from the total number of data.
- Compute center (average) of each cluster.
- Do k-means by using the centers of the clusters of hierarchical as initial centers for its.

Hybrid K-means clustering algorithm (HK) gives good performance and more meaningful results.

Clustering Validation:

Clustering validation is a method to discover a set of clusters that best fits natural partitions (number of clusters) without any a priori class information. There are two types of clustering validation which are based on external criteria and internal criteria. We need to a priori knowledge of data set information when using external validation indexes but it is hard to use in real problems because the real problems do not have prior information of the dataset in question.

When we using internal validity indexes which do not require a priori information from dataset [11].

External validation (Rand Index):

Distance	Ratio	The accuracy of the rand index.		
		Standard k-means	Hybrid Average k-means	Hybrid Single - k-means
Euclidean	0.1	0.8866	0.9959	0.9947
	0.2	0.8856	0.9944	0.9952
	0.3	0.8843	0.9935	0.9939
	0.4	0.8836	0.9925	0.9930
	0.5	0.8830	0.9917	0.9921
	0.6	0.8827	0.8913	0.8913
	0.7	0.8823	0.8849	0.8862
	0.8	0.8821	0.8825	0.8829
	0.9	0.8820	0.8820	0.8821
Manhattan	0.1	0.8892	0.9969	0.9954
	0.2	0.8857	0.9954	0.9963
	0.3	0.8843	0.9945	0.9949
	0.4	0.8835	0.9935	0.9940
	0.5	0.8830	0.9927	0.9931
	0.6	0.8826	0.8923	0.8923
	0.7	0.8824	0.8857	0.8960
	0.8	0.8821	0.8834	0.8839
	0.9	0.8820	0.8829	0.8830

in 1971 Rand proposed an objective criterion for selecting an appropriate clustering algorithm for a given dataset by comparing two clustering algorithms depended on how pairs of data points are clustered [12]. lets $Z=\{z_1,z_2,....z_n\}$ partition into two $S=\{s_1,s_2,....s_r\}$ a partition of Z into r subsets, and $X=\{x_1,x_2,....x_z\}$ a partition of Z into z subsets, define the following:

$$R = \frac{a+b}{a+b+c+d}$$

a, the number of pairs of data points in Z that are in the same subset in S and in the same subset in X.
b,the number of pairs of data points in Z that are in different subsets in S and in different subsets in X.
c, the number of pairs of data points in Z that are in the same subset in S and in different subsets in X.
d, the number of pairs of data points in Z that are in different subsets in S and in the same subsets in X.
- a value of the Rand index between 0 and 1. If the Rand index is less than the expected index then it can give negative values.

6. Compare hybrid hierarchical with standard hierarchical clustering**:**

In this section, two kinds of data, namely real and synthetic data, were used for the evaluation of hybrid hierarchical.

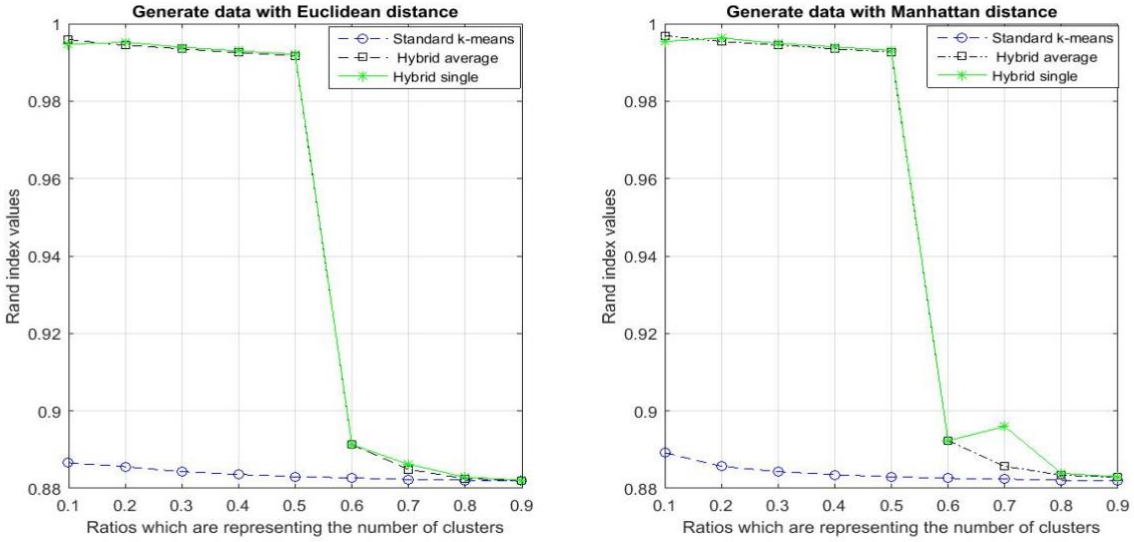
Experiments on Synthetic Data with Normal Distribution n order to evaluate the hybrid algorithm, it is compared with the standard algorithm in term of accuracy on synthetic data (1000 sample*2D) of normal distribution for single and average hierarchical with Euclidean and Manhattan distance.

Table 1. Experiments of hybrid k-means method (HK) on synthetic data of normal distribution

Table 1. represents the experiments on synthetic data of normal distribution for single linkage and average linkage with Euclidean and Manhattan distances. As it is observed in table 1, standard k-means and hybrid k-means methods are compared with each other in terms of accuracy represented by Rand index as one of the external criteria. The results show that the accuracy of the hybrid method is better than that of the traditional method. Also, it is seen in table 1 that there is no difference between two distances for the hybrid k-means method (HK) with normal distribution and also there is no difference between hybrid single linkage and the hybrid average linkage in terms of the accuracy of Rand index. Also, we see the number of

clusters gives good results in terms of the accuracy represented by the Rand index when they equal ratios between 0.1 and 0.6 from the total number of original data.

Figure 1. Compare hybrid k-means (HK) with standard k-means on synthetic data of normal distribution in terms of Rand index.



It is seen in figure1 that there is no clear difference in the Rand accuracy between Euclidean and Manhattan distance and also, there is no difference between the hybrid single linkage and the hybrid average linkage except while the number of clusters is equal to 0.7 where hybrid single linkage has higher accuracy than the hybrid average linkage with Manhattan distance. Generally, the hybrid K- means method (HK) on this data with normal distribution gives the better result than standard K-means in terms of Rand index accuracy.

Discussion the Results of Experiments on Synthetic Data:

The results show the hybrid K- means better than standard K- means in terms of accuracy which is represented by Rand index. The experiments on synthetic data with normal distribution show that there is no difference between the hybrid single linkage and the hybrid average linkage. And also, there is no difference between Euclidean and Manhattan distance in terms of accuracy which is represented by Rand index on the synthetic data of normal distribution. Also, we see the number of clusters gives good results in terms of the accuracy represented by the Rand index when they equal ratios between 0.1 and 0.6 from the total number of original data. As it is clear, the accuracy of the Rand index gradually decreases by increasing the number of clusters represented by percentages from 0.1 to 0.9. The reason is that when the number of clusters is increased for the hybrid method, the similarity to the real data set structure is decreased.

Compare Hybrid K-means with Standard K- means Clustering on Real Data*:*

In order to evaluate the hybrid K-means algorithm (HK), it is compared with the standard K- means algorithm in term of accuracy on the real data (iris dataset) for single and average hierarchical with Manhattan and Euclidean distance.

Fisher's Iris Data Set

in 1936, Ronald Fisher introduced the Iris flower data set in his paper the use of multiple measurements taxonomic problems as an example of linear discriminant analysis. This data set is a multivariate data set and it has three classes where each class refers to a kind of iris flower. These classes are (Iris setosa, Iris virginica, and Iris versicolor) of 50 data points each, this data contains 4 numeric predictive attributes (sepal

length, sepal width, petal length, petal width). Fisher's Iris data set is famous for use databases in pattern recognition. [13].

Table 2. Experiments of hybrid k-means (HK) on Fisher's Iris data.

Distance	Ratio	The accuracy of the rand index.		
		Standard k-means	Hybrid Average k-means	Hybrid Single - k-means
Euclidean	0.1	0.7383	0.7512	0.7381
	0.2	0.7041	0.7145	0.7060
	0.3	0.6912	0.6974	0.6969
	0.4	0.6833	0.6873	0.6877
	0.5	0.6800	0.6840	0.6842
	0.6	0.6777	0.6810	0.6803
	0.7	0.6758	0.6762	0.6778
	0.8	0.6738	0.6745	0.6746
	0.9	0.6724	0.6730	0.6727
Manhattan	0.1	0.7385	0.7572	0.7542
	0.2	0.7020	0.7170	0.7191
	0.3	0.6900	0.6977	0.7072
	0.4	0.6845	0.6906	0.6923
	0.5	0.6800	0.6850	0.6846
	0.6	0.6781	0.6799	0.6805
	0.7	0.6764	0.6770	0.6780
	0.8	0.6743	0.6745	0.6750
	0.9	0.6725	0.6726	0.6729

Table 2 represent the experiments on Fisher's iris data for single linkage and average linkage with Euclidean and Manhattan distances. As it is observed in this table, standard k-means and hybrid k-means methods are compared with each other in terms of accuracy. The results show that the accuracy of the hybrid k-means method is better than the traditional method. Also, it is seen in Table 2 the Manhattan distance gives a better result than the Euclidean distance in terms of an accuracy represented by Rand index, and average linkage method with this distance gives Rand index accuracy better than the single linkage with the same distance. Also, we see the number clusters gives good result in term of the accuracy represented by Rand index when the number clusters equal to ratios between 0.1 and 0.3 from the total number of original data.

Figure 2. Compare hybrid k-means (HK) with standard k-means on iris data set with two distances in terms of the Rand index.

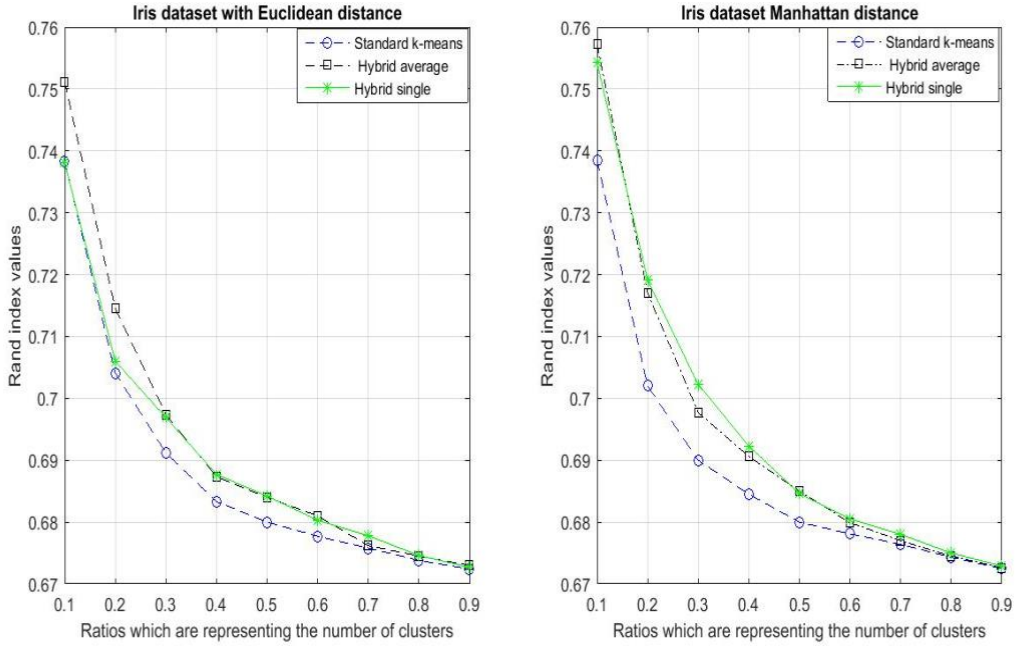


Figure 2 shows comparison between standard k-means and hybrid k-means in terms of accuracy represented by Rand index as one of the external criteria. In figure 2, we found that the average linkage with the Euclidean distance is better than Single linkage with the same distance if the number cluster equal to ratios 0.1 and 0.2. Else it equal to single linkage for the rest of ratios. There is no clear difference between the average linkage and single linkage with the Manhattan distance, except one case when the number of clusters is equal to ratio 0.3 from the total number of original data that the single linkage accuracy is slightly more than average linkage. Generally, Manhattan distance better than Euclidean distance when the number of cluster equal to ratios between 0.1 and 0.4.

Discussion the Results of Experiments on Real Data

In this section, the results of the comparison between hybrid k-means algorithm and standard k-means algorithm are explained on real data:

The results show that the accuracy of the hybrid method is better than the standard k-means method. And also experiments on the iris dataset shows that The Manhattan distance gives a better result than the Euclidean distance for this data the hybrid average linkage with the Euclidean distance is better than hybrid single linkage with the same distance for iris data. Also, we see the number clusters equal to ratios between 0.1 and 0.3 from the total number of original data give good results in terms of accuracy. As it is clear, the accuracy of the Rand index gradually decreases by increasing the number of clusters represented by percentages from 0.1 to 0.9. The reason is that when the number of clusters is increased for the hybrid method, the similarity to the real data set structure is decreased.

Conclusion

In this paper, we used synthetic data which we generated it with normal distribution and real data. Also, we used the single linkage & average linkage and Manhattan distance. Moreover, we determined the number of clusters using ratios in the range between 0.1 and 0.9 from the total number of original data, to evaluate the hybrid K- means method (HK) in term of accuracy which is represented by Rand index (external criteria). The hybrid K- means method HK is proposed to address the problem that initial k (number of clusters) and initial centroids must be selected in prior. In hybrid K- means method (HK), k (number of clusters) and initial centroids was generated via hierarchical clustering methods. Generally, the results show

that the accuracy (external) of the hybrid K- means method (HK) is better than the traditional K- means method.

References

[1] Chen B, Tai P C, Harrison R and et al 2005 Novel hybrid hierarchical-K-means clustering Method (HK-means) for microarray analysis (Computational Systems Bioinformatics Conferences Workshops and Poster Abstracts IEEE) PP105-108.

[2]Ying H xi L Q 2012 Study on PCA based Hierarchical K-means Clustering Algorithm (Journal Control and Automation) PP 6- 068.

[3]Bouguettaya A Yu Q, Liu X , Zhou X and Song A 2014 Efficient agglomerative hierarchical clustering (Expert Systems with Applications) PP 2785–2797

[4]Zhang L and Cui W 2010 Hybrid clustering algorithm based on partitioning and hierarchical method (Computer Engineering and Applications vol 46) PP127-129.

[5]Liu W, Liang Y, Fan J, Feng Z and Cai Y 2016 Improved Hierarchical K-means Clustering Algorithm without Iteration Based on Distance Measurement(International Conference on Intelligent Information) PP38-46.

[6]Bora D J and Gupta A.K 2014 Effect of Different Distance Measures on the Performance of K-Means Algorithm An Experimental Study in Mat lab (International Journal of Computer Science and Information Technologies vol 5) PP 2501-2506.

[7]Rencher A. C 2002 Methods of multivariate analysis (Canada : John Wiley & Sons) chapter 14 PP 451-503.

[8]Shalom SA A and Dash M. 2013 Efficient partitioning based hierarchical agglomerative clustering using graphics accelerators with cuda (International journal of artificial intelligence & applications vol 4) p13.

[9]Celebi, M. E., Kingravi, H. A. and Vela, P. A 2013 A comparative study of efficient initialization methods for the k-means clustering algorithm (Expert Systems with Applications) 40 (1) PP 200–210.

[10] Chow CH , Su M C and Eugene L 2004 A new Validity Measure for Clusters with Different Densities (Pattern Anal. Applications vol 7) PP 2005-2020.

[11] Halkidi M, Batistakis Y and Vazirgiannis M. 2001 On Clustering Validation Techniques (Journal of Intelligent Information Systems the Netherlands) PP 107– 145.

[12] Rand W M 1971 Objective criteria for the evaluation of clustering methods (Journal of the American Statistical Association) PP 846–850.

[13] Fisher, R.A., Iris data set, link = https://archive.ics.edu/ml/datasets/iris , last seen March 3, 2