Data Compression and Hiding using Improved Vector Quantization Techniques

A.Suresh
Assistant Professor
Sri Ramakrishna Institute of Technology
Coimbatore

P. Malathi, Ph.D
Principal
Bharathiyar Institute of Engineering for Women
Salem

ABSTRACT

Generally, Embedding of needful information in a host signal without any loss of host information is required for copyrights protection. Customer identification can be embedded directly into multimedia files, this guide to a number of definite requirements with respect to robustness, simplicity and complexity. In early stages, data hiding can be processed based on vector quantization (VQ) compressed code. Most of this method just hides data into the VQ compressed code, other than not dealing with the quality of VQ decompressed image. Furthermore, if the hiding method is irreversible then the quality of VQ decompressed image shall be not as good as PSNR values and mapping vectors from a vector space to a finite number of regions in that space need more time to search regions. To overcome these problems, vector quantization in data hiding schema proposed a K-Means Multi-Objective Genetic Algorithm (KMOGA) with vector quantization methods. In this proposed approach, a finite number of regions results are found from K-Means multi-objective genetic algorithm and vector quantization technique which can split and compress the data into smaller groups and are embedded into the multimedia pixels. This KMOGA-VQ method enables for effective pixel utilization and retrieves the content without any loss of original pixels. This kind of KMOGA-VQ is based on indices, whereas, the data content can be mapped into many to one form than VQ methods.

Keywords

Data hiding, stenography, water marking, vector quantization, copyright protections.

1. INTRODUCTION

As Internet and digital technology have developed dynamically, people nowadays make use of the Internet extensively to share and transfer digital media including video, image, and music etc. But when people enjoy the expediency of Internet and use these resources extensively, information security issues are produced such as intellectual property protection and infringement of privacy. Make sure the security of digital data transmission has turn out to be a very important issue in recent years and data hiding (Petitcolaset al, 1999) is one of the core technologies. Data hiding can be generally classified into two categories, one for irreversible data hiding, and the other for the reversible data hiding. The previous approach includes several kinds of schemes which demolish the original data after extracting data, such as LSB replacement method (Changet al, 2003; Yang, 2008; Chang and Cheng, 2004) etc., and these methods could not recover image totally after extraction and will result image distortion and damage.

This paper discuss the spreading of digital multimedia has made a copyright Protection a necessary one. Authentication and data hiding become an important issues (Barton, 1997).Watermarking means embedding a piece of information into multimedia content, such as Video, audio or images in that manner it is unnoticeable to a human observer, but easily detected by a computer or detector. Digital watermarking is a message, which is embedded into digital content (images, audio, video or text) that can be detected or extracted later. Such messages typically transmit copyright information of the content. Though, in digital watermarking, the message is not supposed to visible, but electronic devices can retrieve the embedded message to identify the copyright owner. Such a watermark is believed to be robust against distinctive transformation of the content. Therefore the watermark required to be reliably detectable, even is the content (e.g. the image) is cropped, rotated, compressed, etc.

To extract the watermark the same steps of embedding watermark is functioned except of the embedding the extracting operation is performed by subtracting the luminance value of the watermarked image pixel from the luminance of the original image pixel. In this paper, a new clustering based VQ method is proposed for effective pixel utilization and retrieves the content without any loss of original pixels. This kind of VQ is based on indices; the data content can be mapped into many to one form..

2. DATA HIDINGPROPERTIES

2.1 Robustness

A watermarking scheme should be in opposition to damage from standard image processing and malicious attacks. For instance, watermarked images possibly will be compressed before transmitting or accumulate it. Hence, the watermarked image has to endure the suitable practice such as lossy compressions, conversions, re-samples and other non-malicious operations. Alternatively, intentional or the unintentional attacks are attempt to remove the embedded watermark (Bruyndonckx et al, 1995). A robust watermarking scheme has to make sure the retrieved watermark is renowned, when the image quality does not obtain seriously damaged.

2.2 Imperceptibility

A watermark can be embedded into an image as either noticeable or indiscernible. The visible watermark mostly removed by a noise removing process. So as to decrease the hazard of cracking, most of the proposed watermarking schemes are invisible (Jayaraman et al, 2009). If the embedding process critically affects the quality of the watermarked image, in addition the watermarked mark typically affected by attackers or loses its pixel values. Consequently, the quality between the original image and the watermarked image is supposed to not be seriously corrupted. The property is called imperceptibility.

2.3 Reversible Watermarking

The overheads of embedding process and retrieving process should be restricted in a reasonable range is called reversible watermarking. Distinct to conventional watermarking schemes, reversible watermarking schemes make the cover image without loss of original image. This show the nearby pixel value will not be affected much. The affected value are very less compare than conventional watermarking.

3. EXISTING WORKS FOR COPYRIGHT

3.1 Spatial-domain and frequency-domain techniques

As a general rule, any watermarking technique is made either in Spatial or Frequency domain. Spatial Domain techniques are particularly easy to build and design, and then they provide a perfect reconstruction in the absence of noise as demonstrated in the results. There are numerous techniques are there in spatial domain embedding utilizing the luminance components (Jamal Hussein, 2010), manipulating the Least Significant Bits (Nikolaidis and Pitas, 1998) as ideal locations for embedding. However, the watermark can be simply damaged if the watermarked image is low-pass filtered or JPEG compressed. Here, a pseudo-random number generator is used to find out the sequence of locations on the image plane. Manipulating the Intensity Components (Verma et al, 2007), Image Differencing (Wu and Tsai, 2000) etc are the other approaches presented.

In contrast, in frequency domain, the cover image and the watermarks are supposed to a transformation into the frequency domain where deeper manipulations of the coefficients are possible without visible degradation to the cover image is possible. The transformation may be processed through Discrete Cosine Transform (Juan Hernandez et al, 2000), Discrete Wavelet Transform (Ali Al-Haj, 2007), Ridgelet Transform etc. The watermark is then embedded in the transformed coefficients of the image such that the watermark is less invisible and more robust to some image processing operations. Finally, the coefficients are inverse-transformed to form the watermarked image.

3.2 Difference Expansion

The DE embedding technique engages into pairing the pixels of the host image and transform of a pair of pixels to work out a high capacity and low distortion reversible watermark. Based on the difference between the original image and reconstructed image this technique will produce high PSNR difference, which will lead more robustness.

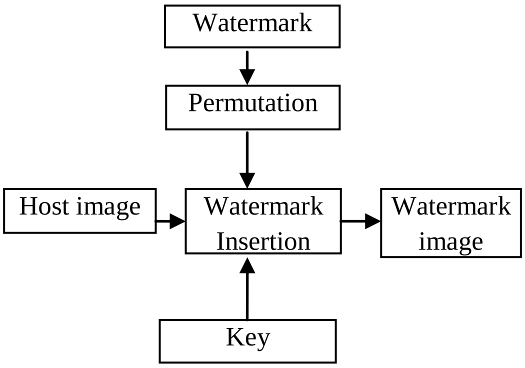

Fig 1: Block diagram of the watermarking Insertion system

3.3 Watermark Embedding methods

In this approach, the embedded watermark has to be invisible to human eyes and simple to the majority of image processing operations. For that, a bit of binary pixel value (0 or 1) is embedded in a block of the host image. Before insertion, the host image is decomposed into NxN blocks. Based on the difference of a block, pixels in the block are adaptively customized to maximize robustness and assure invisibility.

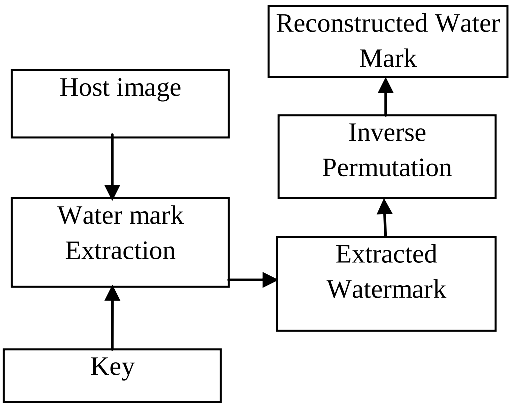

Fig 2: Block diagram of the watermarking Extraction system

The place or block for embedding is chosen by a pseudo-random number generator by means of a seed value k

$$g_{max} = \max(b_{ij}, 0 \leq i, j < N)$$ and

$$g_{max} = \min(b_{ij}, 0 \leq i, j < N)$$

Where b_{ij} represents the intensity of the (i, j)-th pixel in block B. Let assume that the embedded pixel value b_w is 0 or 1.The embedding procedure adjust or modifies the intensities of pixels in the block B according to the following rules

1. $b_w = 1$

$$g' = g_{max} \quad \text{if } g > g_{mean}$$

$$g' = g + \delta \qquad \text{if } g \leq g_{mean}$$

2. $b_w = 0;$
$g' = g_{min}$ if $g < g_{mean}$
$g' = g - \delta$ if $g \geq g_{mean}$

Where g' is the modified intensity and d is a small value used to adjust the intensities. The embedding of the watermark depends on the content of each block. If the block is of higher contrast, the intensities of pixels will be modified very much. Or else, the intensities are tuned somewhat Therefore the proposed algorithm can adaptively modify the content of a block. Let blocks B and B' denote the original and watermarked blocks, respectively. The sum of pixel intensities of B' will be larger than that of B if the inserted watermark pixel value b_w is 1. In contrast, if the inserted watermark pixel value b_w is 0, the sum of pixel intensities of B' will be smaller than that of B (Lu and Sun, 2000).

3.4 Watermark Extractions methods

The extraction of a watermark is similar to the embedding process whereas in a reverse order. In this algorithm, the extraction of a watermark should make reference to the original host image. Initially, make use of the seed value, k, to produce a sequence of positions or blocks where the watermark is embedded. For each selected position, let B and B' correspond to the analogous blocks of the original host image and watermarked image, correspondingly. Calculate the sum of pixel intensities, S_0 and S_w, of B and B'. The retrieved watermark bit value b_w is determined by the following rule:

$b_w = 1$ if $S_w > S_0$,

$b_w = 0$ if S_w£ S_0

The extracted watermark bit values, b_w's, are then inversely permutated to get the reconstructed watermark.

Almost, image frequently requires some compression techniques for the duration of transmission on internet, such as JPEG and VQ, etc., but if an image with embedded data is compressed, it may result the hided data being destroyed. To deal with this problem, several researchers were talk about data hiding schemes designed on VQ compressed codes (i.e., the index table) (Jo and kim, 2002; Chang et al, 2007; Yang and Lin, 2009).

4. VECTOR QUANTIZATION AND IMPROVED K-MEANS MULTIOBJECTIVE GENETIC ALGORITHM (KMOGA) BASED VECTOR QUANTIZATION (KMOGA-VQ)

4.1 Vector quantization

Vector quantization permits the modeling of probability density functions by the distribution of archetype vectors. It is the process of splitting up of a large set of points (vectors) into groups having about the same number of points nearer to them. Each group is represented by its centroid point, as in k-means and several other clustering algorithms. The density matching property of vector quantization is dominant, particularly for finding the density of large and high-dimensioned data. As data points are correspond to the index of their neighboring centroid, generally taking place the data having low error and rare data high error. This is why VQ is appropriate for lossy data compression. It can also be used for lossy data correction and density estimation.

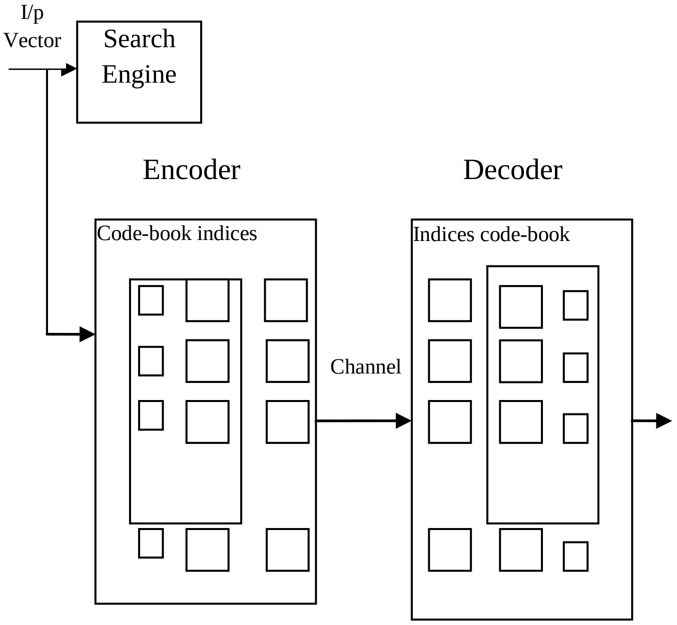

Fig 4: Vector quantization Scheme

Before going into vector quantization, should have to know about two techniques.

1. Mapping technique (grouping the values)
2. Designing a code book (mechanism of mapping or entries of code words)

The number of code vectors (N) depends upon two parameters, they are, rate (R) and dimensions (L). The n code vectors are considered through the following formulae:

Number of code vectors (N) =*2R•L.* where

R — Rate (bits/pixel)
L -- Dimensions (grouping)
When the rate increases, the number of code vector increases [20]. As the number of code vectors increase. Size of the code book also increases.

You are given an image of size X(m, n) =

$$\begin{vmatrix} 02 & 04 & 06 & 08 \\ 10 & 11 & 16 & 15 \\ 09 & 03 & 01 & 07 \\ 12 & 14 & 13 & 05 \end{vmatrix}$$

From the input image, the following data can be obtained through visual inspection:

1. The maximum value in the input image is sixteen.
2. The minimum value in the input image is one.

4.2 Improved K-Means Multi-Objective Genetic Algorithm (KMOGA) based vector quantization (KMOGA-VQ)

Vector quantization (VQ) is a process of mapping vectors from a vector space to a finite number of regions in that space adapts more time to search the pixel values or regions to embed the data. To overcome the problem of the Vector Quantization, a K-Means Multi-Objective Genetic Algorithm is proposed. Whereas, the Multi-objective genetic algorithm with Pareto rank approach can be used to increase the K-means performance. This approach capitulate a set of solution that comprises of several fronts based on their ranks. The first Pareto front comprises of non-dominated solution, whereas it consists of a pair of values where the distance between points in a cluster is minimum and the inter-cluster distance between clusters is maximum. The minimum Davies-Bouldin validity index equation (9) and appropriate cluster number are employed to locate the optimal solution. These regions are called as clusters and correspond to their central vectors or centroids. A set of centroids, which deals with the whole vector space, is known as a codebook.

VQ is a mapping function which maps k-dimensional vector space to a finite set $CB = \{C1, C2, C3, \dots\dots, CN\}$. The set CB is called as codebook comprises of N number of codevectors and each codevector $C_i = \{c_{i1}, c_{i2}, c_{i3}, \dots\dots, c_{ik}\}$ is of dimension k. High-quality codebook design leads to reduced distortion in reconstructed image.For encoding, image is split in blocks and each block is then converted to the training vector $X_i = (x_{i1}, x_{i2}, \dots\dots, x_{ik})$. The stages of K-Means clustering mainly comprises of initializing the population, making objective function, counting fitness function based on pareto ranking, crossover, mutation, etilism until the criteria achieved.

4.2.1. Initializing population

The initial population is through by formative the length of chromosome with size$K \times d$. K is the number of chromosome with numerous d, while d is the dimension of the cluster variables. $Vd_1, \dots Vd_k$ is a random value which generated based on the minimum and maximum values of the variable d.

In data hiding process the VQ can be used to compress an image both in the spatial domain and in the frequency domain. The data codebook is generated.

4.2.2. Objective Function

K-Means clustering optimization with multi-objective genetic algorithm uses two objective function to select the most appropriate pixel search space objectives, i.e. minimizing variance functions within each cluster (eq 1) and maximizing the variance between cluster (eq 2). The calculations used as follows,

$$\sigma_i^2 = \frac{1}{N}\sum_{j=1}^{N}\sigma_{ij}^2 \rightarrow (1)$$

Where $i = 1,2,\dots\dots\dots k$ K is the number of clusters

$$\sigma^2 = \sum_{i=1}^{K}\sigma_i^2 \rightarrow (2)$$

Whereas

σ_i^2:variant in the i^{th} cluster

n_i:number of data in the i^{th} cluster

x_{ij}: Data in i-cluster ,j^{th} variable

Z_{ij}: i-cluster average in j^{th} variable

v:Number of variable

The initial function is to minimize variant average in cluster which formulated as follow:

$$V(w) = \frac{1}{k}\sum_{i=1}^{k}\sigma^2 i\,, i = 1,2,\dots k \rightarrow (3)$$

Whereas

$V(w)$: Variant in cluster

K : Number of cluster

$$V(b) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{v}(Z_{ij} - \bar{Z}_j)\,, i = 1,2,\dots k \rightarrow (3)$$

Where

$V(b)$: Variant in cluster

Z_{ij}: i-cluster average in ,j^{th} variable

$\bar{Z}_j$: Grand mean of j^{th} variable

Fitness function of the centroid value is calculated by pareto ranking approach, where each individual datum is calculated by the overall population based on the non-domination concept. After that, pareto ranking approach is processed by equations,

$$r_2(x,t) = 1 + nq(x,t) \rightarrow (4)$$

Whereas

$r_2(x,t)$:x^{th} completion rank in the t^{th} iteration

$nq(x,t)$:Solution number which dominate x completion in the t^{th} iteration

The crossover is a genetic operator in generating new chromosome (offspring). In this learning, single point crossover is used with crossover probability which is calculated by the following equation

$$pz = \frac{F(S_z)}{\sum_{z=1}^{Z}F(S_z)} \rightarrow (5)$$

Whereas

pz: z^{th}completion selection probability

$F(S_z)$:fitness value in S_z solution

4.2.3. Mutation

Another genetic operator is a mutation process. This process is utilized against the possibility of modifications to the existing results.The selection process of chromosome mutations, in addition to the position of a gene to be transformed will be made in a random. The number of affected offspring is determined by the mutation probability as in equation

$$p_k = \frac{1.5 * d_{max}(X_n) - d(X_n, c_k) + 0.5}{\sum_{k=1}^{K}(1.5 * d_{max}(X_n) - d(X_n, c_k) + 0.5)} \rightarrow (6)$$

Whereas

p_k mutation probability of code vector in k^{th} cluster

$d(X_n, c_k)$:jarak euclidean distance between X_n data code vector and c_k center point of k^{tn} cluster

$d_{max}(X_n) : \max_k\{d(X_n, c_k)\}$

4.2.4. Elitism

The random selection doesn't assure the non-dominated solutions will stay alive in the next generation. The real step for elitism in multi-objective genetic algorithm is doubling-up the non-dominated solution in population P_t, henceforth will be included in the population of P_{t+1} by selecting the non-dominated solution.

4.2.5. Davies-Bouldin Index

Davies-Bouldin index is worn to maximize the distance among the cluster $C_i$and C_j, and certainly, it is used to minimize the distance between the points in a cluster with the center of the cluster. The distance $s_c(Q_k)$ within a cluster Q_k is defined by:

$$s_c(Q_k) = \sum_i \frac{||x_i - c_k||}{N_k} \rightarrow (7)$$

Where N_k is the number of points that belong to the codebook cluster Q_k and

$$C_k = \frac{1}{N_k} \sum X_i \rightarrow (8)$$

The distance beween the clusters is defined as

$d_{cc} = ||C_k - C_l||$ so that the DB index DB index is defined as

$$DB\,(nc) = \frac{1}{nc} \sum_{k=1}^{nc} \max_{l \neq k} \left\{ \frac{s_c(Q_k) + s_c(Q_i)}{d_{cc}(Q_k, Q_l)} \right\} \rightarrow (9)$$

Algorithm 1:

$\mathrm{K - Means\ Multi\ Objective\ Genetic\ \ Algorithm(KMOGA) - Ve}$

Step 1: Divide the image into non overlapping blocks and adapt each block to vectors therefore forming a training vector set. For clustering a set of L training vectors into a set of M codebook vectors.

Step 2: initialize $i = 1$;

Step 3: Compute the centroid (codevector) of this training vector set.

Step 3.1 K-Means Multi-Objective Genetic Algorithm (KMOGA) for each training vector, find the codeword in the current codebook that is closest and assign that vector to the equivalent cell connected with the closest codeword from equation (9).

Step 4: Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

Step 5: Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold

Step 6 : Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed

5. RESULTS

Fig 5: Cameraman Image

For analyzing resultant 2 mb Data content that is to be split using vector quantization method by 32 bit data chunks. This data chunks will be embedded into cameraman image for copyright protection. This hidden data can be retrieved by using DES encryption and decryption method, Difference of normal image and the encrypted image can be proven by PSNR values in table below

Table 5: PSNR Value

Parameters	**Payload Before Embedded data**	**Payload After Embedded data**
MSE	1.565	1.601
PSNR	21.5 dB	21.47

6. CONCLUSION

In this paper, mainly concentrate on the minimum loss of payload when reconstructed data at final stage. Initially some data like 2 mb, 4mb, and 6mb are taken for the experiment. This data should be dividing by means of the data compression technique called Improved Vector Quantization with K-Means Multi-Objective Genetic Algorithm (KMOGA). Using this vector can be group the data centroid means minimum Davies-Bouldin validity index. Through which it, improves the effective utilization of the data at mid of pixels and also produce smooth reconstructed image at less loss of payload, this method will split the data by 32 bit each, so the hiding data to another image or text would be very easier. These results can be concluded that K-Mean GA with VQ could reach a better optimal solution than K-Mean, which is able to find minimum index value for each code vector. Using the paper can send the data from one location to another location without further intervening. The hidden data can be viewed only by the respective receiver by using corresponding password, but this paper proposed to a approach to splitting and hiding, apart from this DES encryption is involved for providing the encryption and decryption operation. This method is useful for effective compression and safe sending information. The PSNR ratio also is found for further analysis.

7. REFERENCES

[1] Petitcolas, FAP, Anderson, RJ and Kuhn, MG, 1999. “Information hiding-a survey,” Proceedings of the IEEE, Vol.87, No. 7, pp. 1062-1078.

[2] Chang, CC, Hsiao, JY, and Chan, CS, 2003.“Finding Optimal Least Significant-Bit Substitution in Image Hiding by Dynamic Programming Strategy,” Pattern Recognition, Vol. 36, No. 7, pp. 1583-1595.

[3] Yang, CH, 2008.“Inverted Pattern Approach to Improve Image Quality of Information Hiding by LSB Substitution,” Pattern Recognition, Vol. 41, No. 8, pp. 2674-2683.

[4] Chang, CK and Cheng, LM, 2004. “Hiding Data in Images by Simple LSB Substitution,” Pattern Recognition, Vol. 37, No. 3, pp. 469-474.

[5] Barton, JM, 1997. “Method and apparatus for embedding authentication information within digital data,” U.S. Patent 5 646 997.

[6] Bruyndonckx, O, Quisquater, JJ and Macq, B, 1995. "Spatial method for copyright labeling of digital images", Proceedings of IEEE Nonlinear Signal Processing Workshop, pp. 456-459.

[7] Jayaraman S, Esakkirajan S, Veerakumar T, 2009. “Digital image processing copyright”,tata Mcgraw Hill Education private limited.

[8] Jamal Hussein, 2010. “Spatial domain Watermarking scheme for color images based on log average luminance”, Int’l Journal of Computing, Volume 2, Issue 1.

[9] Nikolaidis and Pitas, 1998. “Robust Image Watermarking the Spatial Domain”, International Journal of Signal Processing”, Vol.66, Issue 3, pp. 385 – 403.

[10] Verma P, Agarwal, DP and Jain S, 2007. “Spatial Domain Robust Blind Watermarking for Color Image”, Asian Journal of Information Technology, Vol. 6, Issue. 4, pp. 430 – 435.

[11] Wu, DC, Tsai, WH., 2000. “Spatial-Domain Image Hiding Using an Image Differencing”, IEEE Proceedings -Vision, Image and Signal Processing, Vol. 147, Issue. 1, pp. 29 – 37.

[12] Juan Hernandez, Martin Amado, Fernando Perez, 2000. “DCT Domain Watermarking Techniques for Still Images: Detector Performance analysis and new structure”, IEEE Transactions on Image Processing, Vol. 9, No.1, pp. 55 – 68.

[13] Ali Al-Haj, 2007. “Combined DWT – DCT Digital Image Watermarking”, Journal of ComputerScience, Vol.3, Issue.9, pp.740-746.

[14] Lu, ZM, Sun, SH, 2000. “Digital image watermarking technique based on vector quantisation," Electronics Letters, Vol. 36, No. 4, pp. 303-305.

[15] Jo, M and Kim, HD, 2002. “A digital image watermarking scheme based on vector quantization,” IEICE Transactions on Information and Systems, Vol. E85-D, No. 6, pp. 1054-1056.

[16] Chang, CC, Wu, WC and Hu, YC, 2007. "Lossless recovery of a VQ index table with embedded secret data," Journal of Visual Communication and Image Representation, Vol. 18, No. 3, pp. 207-216.

[17] Yang, CH and Lin, YC, 2009.“Reversible data hiding of a VQ index table based on referred counts,” Journal of Visual Communication and Image Representation, Vol. 20, No. 6, pp. 399-407.

A. SURESH completed B.E electrical and electronics engineering degree programme at Government College of Technology (GCT) and M.E computer science and engineering at Sri Krishna College of engineering and technology. He is pursuing Phd under Anna University Chennai. He is currently working as an Assistant Professor in the department of Information Technology at Sri Ramakrishna institute of technology, Coimbatore. His area of Interest includes Soft Computing, Neural Networks and Digital Image Processing.

Dr.P. Malathi has received her M.E degree in Applied Electronics from Coimbatore Institute of Technology, Coimbatore in 2001. She completed her Ph.D at PSG College of Technology, Coimbatore. She has the teaching experience of 15 years. She is the member of IEEE, ISTE and IETE. Her areas of interest are OFDM, MIMO, WLAN, PLC and Wireless Mobile Communication. Presently, she is the Principal ofBharathiyar Institute of Engineering for Women, Salem.