Ann. For. Res. 66(1): 849-869, 2023
ISSN: 18448135, 20652445

ANNALS OF FOREST RESEARCH
www.e-afr.org

DEEP LEARNING-BASED WAVELET EMBEDDING FOR COVERT AUDIO OBJECT EMBEDDING IN VIDEO OBJECT STEGANOGRAPHY

Alaknanda S. Patil [1] and Dr. G. Sundari [2]

[1] *Research Scholar, Department of EC Engineering, Sathyabama Institute of science and technology, Chennai, India*

[2] *Professor, Department of EC Engineering, Sathyabama Institute of science and technology, Chennai, India*

[1]patilalaknanda@yahoo.com

Abstract

Video watermarking relates to the action of embedding a secret text, audio or file into another video. Generally, the video watermarking is carried out based on two steps, like embedding and extraction phase. Embedding phase embeds the secret information on the video using embedding algorithm, whereas the extraction phase retrieves the hidden data on the embedded video using extraction algorithm. In this research, the video object watermarking is done based on embedding and extraction phase. Here, embedding phase embeds the secret audio message on the object location of video, such that the object location is determined by 'Shepard convolution neural network' ('ShCNN'). In the extraction process, the confidential image is extracted from an immerged video by applying extraction algorithm. Moreover, the embedding of audio signal is hard to process; hence to simplify the process, the audio signal is converted into binary format using Quantum Representation of Digital Audio (QRDA) technique. Furthermore, the training procedure of 'ShCNN' is completed utilizing an optimization technique, termed Improved Invasive Honey Badger Optimization (IIHBO) algorithm. Furthermore, an investigational findings provides that the developed plan produced superior results according to the Mean square error Correlation coefficient, and Peak signal to Noise Ratio of 0.028, 0.900 and 46.19, appropriately.

Keywords: Shepard convolution neural network, Quantum Representation of Digital Audio, Honey Badger Optimization, Improved Invasive weeds optimization, Active contour segmentation.

1. Introduction

The digital technology-driven information revolution to present period is large [16], which is deliberated as a challenging task. Some of the advanced multimedia devices, such as powerful camcorder, digital camera and digital voice recorder provide the huge multimedia information [17]. Nowadays, various modalities of online digital information are increased, which improves the communication as faster and easier. Video watermarking is the process of hiding a secret information, image or audio on another video. The advantage of watermarking is that it ensures the ownership, integrity, authentication and copyright protection, which is more challenging. Digital steganography is also same like as watermarking, which is utilized to handle the large issues of embedding. The term steganography refers to the hidden information communication,

such that the secret information is broadcasted by obscuring it on cover media, there by the transmitted secret data is not recognizable. Generally, the cover media as well as secret information can be in any manner, such as video, audio, message or image. Majority of the digital steganography approaches concentrate only on images and texts [5]. Preserving video from unlicensed access is significant in various applications. For instance, the various social networking as well as file sharing applications, like Twitter, You Tube as well as Face book [15] [5].

The primary motive of this research is to embed the secret audio on the predicted region of video frame. For that, the frame is extracted and the object is extracted using active contour segmentation. An exact location of extracted object is tracked by the 'ShCNN', which has been trained by the devised IIHBO technique. Moreover, the embedding process is done by embedding the binary form of secret audio on the combination of wavelet transform and predicted region. In the stage of retrieval, the hidden object is retrieved by applying the similar steps of embedding algorithm on the embedded video, but only difference is that instead of applying QRDA embedding, QRDA retrieval algorithm is applied to retrieve the object.

The primary significance of this investigation is,

- ***Recommended IIHBO-based 'ShCNN' for object region prediction***: Here, the object region is anticipated utilizing 'ShCNN' in which the weights of 'ShCNN' are trained by IIHBO. The IIHBO is designed by the assimilation of 'Improved Invasive Weed Optimization' (IIWO) and Honey Badger Optimization (HBO). Moreover, the embedding is done by applying the QRDA embedding algorithm and the retrieval of hidden video is done by applying QRDA extraction algorithm. Moreover, QRDA algorithm transforms the audio signal into the binary format for simplifying the embedding process.

This article is organized as follow. Portion 2 describes the examination of the research of multimedia object watermarking, the proposed embedding with extraction algorithm is portrayed in portion 3, portion 4 describes the description of the findings and portion 5 presents this study's conclusion.

2 Literature Survey

This section elaborates a review of the research on numerous current video steganography techniques. Minkyung Kwak and Youngho Cho [1] developed the bot-net model for performing video steganography in telegram messenger. This method realized the better performance in terms of communication efficiency. However, this method failed to curtail the data loss of embedded message; thereby it affects the system performance. For progressing the system performance, Yuanzhi Yao and Nenghai Yu [2] modelled the Payload allocation strategy for analysing the modification distortion in video steganography. This method attained the lower computational complexity, but this approach ignored novel deep learning classifiers in order to achieve better achievement. Moreover, it did not provide accurate findings with complicated environments. For attaining the best achievement with difficult environments, Mukesh Dalal and Mamta Juneja [3] devised an effective video steganography technique using Discrete Wavelet Transformation (DWT). In this strategy, the hidden message was placed in the sub band's mid frequency area.

Moreover, this method was more robust against various types of noise attacks, but it is unsuccessful to progress the performance security. For achieving the better security, Hongguozhao *et al*. [4] developed transform block decision for performing the video steganography. This method attained the enhanced visual quality as well as larger embedding capacity, however it was vulnerable to several kinds of powerful attacks.

In order attain the better embedding result with harmful attacks, Tanveer J. Siddiqui and Ashish Khare [5] modelled the Chaos-based Video Steganography scheme with DCT for generating the stego video. The hidden information was incorporated in this method on the middle frequency region of DCT coefficient. Although, this method improved the system security, it was failed to minimize the computational time. To shorten the calculation time, Ke Niu *et al*. [6] devised a Hybrid Adaptive embedding approach for multimedia steganography. Here, the secret information was immerged on the video by combining the probability of embedding and the cost function of distortion. This method improved the execution speed of the algorithm. It was failed to process this method for audio and images to improve the performance outcomes. In order to embed the audio and images on video, Ramadhan J. Mstafa *et al*. [7] devised the secure video steganographic algorithm by combining the DCT, DWT and Multiple Object Tracking (MOT). In this approach, the secret information was encoded with pre-processing approach before embedding. This method improved the embedding capacity and imperceptibility, but this method failed to progress the visual quality as a way to achieve precise results. To be able to recover the visual excellence of embedded video, Rachna Patel *et al.* [8] devised the Discrete sine transform-Secret place of bit for Message Non-dynamic Region (DST-SBPNRM) for video steganography. This approach effectively minimized the complexity level of embedding process. However, it did not utilize the large count of covering multimedia frames to hide excess bits of confidential information by sustaining the rate of concealing as well as bit error rate.

3. Proposed IIHBO-based 'ShCNN' for video object watermarking

Generally, watermarking is utilized to conserve the secret information and to represent the legitimacy of permissible documents. Video watermarking is an embedding approach, which is used to hide the text or logo on video. In this research, the video object watermarking is carried out on two phases, namely embedding phase as well as extraction stage. In the immerging stage, the mystery audio information is fed on the input video in order to get the embedded video. In the stage of retrieval, the mystery audio data is extracted by performing the reverse operation of embedding phase.

Phase 1: Embedding stage

In the immerging stage, the key frame is extracted from input video, and then the object is extracted from the key frame using active contour model. After that, the object region is predicted from the extracted object for embedding the secret message utilizing 'ShCNN' where the weight of 'ShCNN' is given training by the devised IIHBO algorithm. Moreover, the IIHBO is modelled by joining the IIWO and HBA. On the other hand, from the extracted key frame, the wavelet

transform is carried out to partition the key frames into the sub bands. In this research, the audio message is considered as a secret data. In order to perform embedding, the time and amplitude of secret audio message is binary transformated by applying the QRDA scheme. Lastly, the embedding process is done by applying the binary transformation of sound message on the combined region of wavelet transform and the predicted object region. Afterwards, the reverse of wavelet transform is subjected to retrieve the immerged video. Figure 1 shows the embedding phase of devised model.

Let us assume, the database P comprises s numbers of videos, which is mathematically expressed as,

$$P = \{R_1, R_{2,...,} R_v, ..., R_a\} \quad (1)$$

where, a indicates the total count of videos, and R_v signifies the v^{th} video from dataset. In this research, the v^{th} video is deliberated as an input of key frame extraction phase

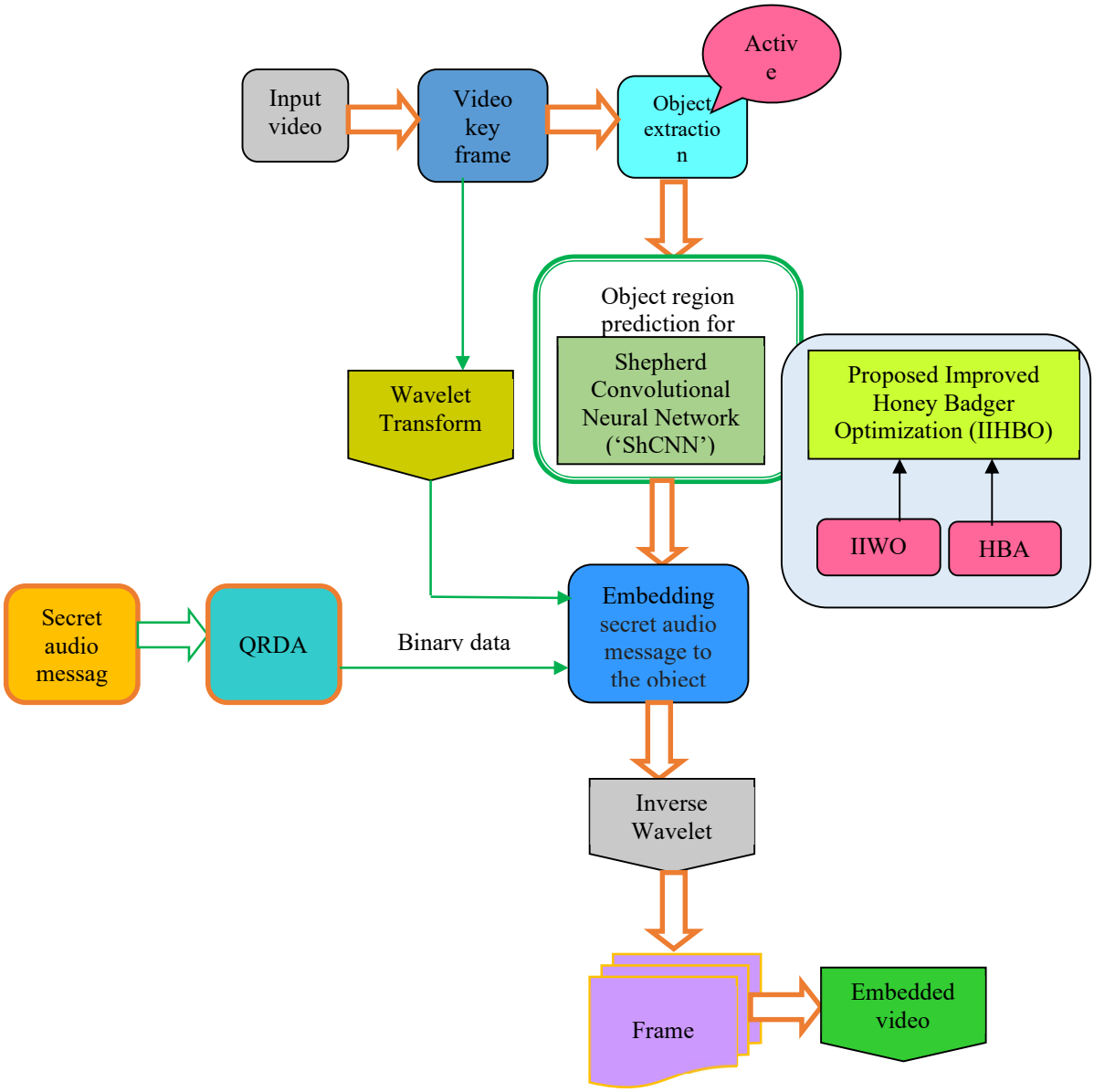


Figure 1. Block diagram of proposed embedding phase

.

3.1 Video frame extraction

It is the process of partitioning the video into multiple frames. In the video frame extraction method, the video R_v is regarded as an input for further parsing. The importance of pre- parsing is to improve the system performance by reducing the data to be processed. The number of extracted frame depends on the frames per second of a processed video. Thus, the long duration videos have large number of frames, whereas the short duration of video has less numbers of frames. Then, the number of frame is calculated by,

$$f_n = fps \times t \tag{2}$$

where, the number of frames are indicated as f_n, fps signifies the frames per second, and t indicates the video duration. In this research, the extraction of key frame is done by Discretetchebichef transform (DTT)and Structural Similarity Index Measure (SSIM). DTT [18] is a transformation technology, which considers the Chebyshev polynomials for its operation. DTT is widely utilized in various image processing applications. Let us assume $D(n)$ be the DTT of an input video $U(a)$, and is expressed as,

$$D(n) = \sum_{k=0}^{K-1} h(n,k)U(a) \tag{3}$$

where, $h(n,k)$ indicates the orthogonal of discrete Chebyshev polynomial. Likewise, the inverse of DTT is expressed as,

$$U(a) = \sum_{k=0}^{K-1} h(n,k)D(n) \tag{4}$$

where, $h(n,k)$ characterize the orthogonal basis of discrete Chebyshev polynomial. After the completion of DTT, the similarity among two frames are needed to be calculated using SSIM. SSIM is a matching technique, which is used to compute the similarity among two consecutive frames. Moreover, the value of SSIM is either 0 or 1, such that the value of '1' signifies the identical frames, whereas '0' signifies non-identical frames. SSIM selects the relevant frames and removes the irrelevant frames by comparing the two frames. The attained key frame is a gathering of various frames, which are distinct to each other. Assume x and y specifies two windows with common size $L \times L$ and is portrayed as,

$$SSIM = \frac{\left(2\alpha_x\alpha_y + \gamma_1\right)\left(2\beta_x\beta_y + \gamma_2\right)}{\left(\alpha_x^2 + \alpha_y^2 + \gamma_1\right)\left(\beta_x^2 + \beta_y^2 + \gamma_2\right)} \tag{5}$$

where, α_x characterizes the estimate value of x, α_y indicates the estimate value of y, β_x^2 denotes the deviation of x, β_y^2 specifies the deviation of y, $\beta_{x,y}$ denotes co- deviation of x and y, γ_1 and γ_2 are the constants. Thus, the extracted key frame is specified as,

$$R = \left\{T_1, T_{2,\ldots,} T_v, \ldots, T_a\right\} \tag{6}$$

where, a signifies the overall frame count, T_v denotes the v^{th} frame from video. After the completion of key frame extraction, the object extraction is carried out using active contour region segmentation model, which is explained as below.

3.2 Object extraction using contour region segmentation

The input of object extraction is considered as T_v, which is utilized to excerpt the object from the frame using active contour region segmentation model [19]. Generally, the image with homogeneous region offers a small difference in intensity, whereas the image with inhomogeneous region offers a large difference in intensity. Thus, the difference among intensity is considered as an important feature for separating the foreground region from thewhole frame. The advantage of active contour region segmentation is that it assumes less amount of energy. In this research, the active contour region segmentation scheme is utilized to extract the particular object from the frames. In order to extract the object from frame, initially, the video frame is categorized into foreground and background area, and the brightness among these two regions is independent. In this method, the outline is formed undiversified in an unsupervised way to diminish the function of energy, such that the lowest energy is attained when the outline extents the border among foreground as well as background areas. The energy function of this model is portrayed as,

$$\varepsilon.Length(G) + \psi.Area(G_{in}) + \lambda^{+}\int_{G_{in}}\left(F(c) - G^{+}(G)\right)^2 + \lambda^{-}\int_{G_{out}}\left(F(c) - G^{-}(G)\right)^2 \tag{7}$$

where, G denotes the contour, $F(c) \in N$ indicates the intensity of image at location c of pixel in spatial domain Ξ, $\varepsilon \geq 0$ represents the standardize constraint, which manages the contour smoothness, G_{in} denotes the forefront region, G_{out} depicts the background region and $\psi \geq 0$ specifies the regularization constraint, which disciplines a huge area of foreground. Moreover, $G^{+}(G) = average\left(F(c)\middle| c \in G_{in}\right)$ and $G^{-}(G) = average\left(F(c)\middle| c \in G_{out}\right)$ indicates the mean intensities of foreground as well as background, correspondingly. The constraint $\lambda^{+}, \lambda^{-} \geq 0$ manages the impact of two energy term of image $\lambda^{+}\int_{G_{in}}\left(F(c) - G^{+}(G)\right)^2$ and $\lambda^{-}\int_{G_{out}}\left(F(c) - G^{-}(G)\right)^2$ at both inside as well as outside of contour. Moreover, the contour G is portrayed as zero level, and is given by,

$$G := \left\{c \in \Xi : \Phi(c) = 0\right\} \tag{8}$$

From the above expression, the separation of foreground as well as background contour G is done by,

$$G_{in} := \left\{c \in \Xi : \Phi(c) > 0\right\} \text{and } G_{out} := \left\{c \in \Xi : \Phi(c) < 0\right\} \tag{9}$$

The foreground region of the segmented frame is processed for detecting the object region. Thus, the extracted object is specified as Z_d, which is fed to the input of object region prediction.

3.3 Object region prediction using 'ShCNN' for embedding

The object region prediction is done with 'ShCNN' by selecting Z_d as an input. The gain of utilizing 'ShCNN' is that the parsing efficiency of this classifier is high. In this research, the 'ShCNN' predicts the region of object exist in the frame, which enhances the effectiveness of video object embedding. The 'ShCNN' is the modified version of CNN, which is formed by adding the shepard interpolation layer on the ReLU layer of CNN. Furthermore, the framework of 'ShCNN' is briefly elucidated in the succeeding portion.

3.3.1 'ShCNN'

'ShCNN' [9] is a neural network, which is entirely same as CNN excluding the ReLU layer. In 'ShCNN', the shepard interpolation layer is included instead of ReLU layer. The convolution format of Shepard model is expressed as,

$$I_p = \begin{cases} (E * O)_p (E * X)_g & ; if\ X_p = 0 \\ O^p & ;\ if\ X_p = 1 \end{cases} \tag{10}$$

where, O signifies the input, I denotes the output, image coordinates are indicated as g, X characterizes the binary indicator, X_p specifies the values fun predicted pixel, $*$ signifies the convolution operation, and E portrays the kernel function. In addition, the component-wise division handles the information, which is broadcasted over the network. Moreover, the 'ShCNN' have been able to handle the interpolation of discontinuous-spaced information. The 'ShCNN' offers the effective prediction outcome through the inclusion of Shepard interpolation layer.

- ***Layer of Shepard interpolation:-*** The expression for layer of interpolation is represented as,

$$A_g^j\left(A^{j-1}, X^j\right) = \omega\left(\sum_q \frac{E_{gr}^j * A_r^{j-1}}{E_{gr}^j * X_r^j} + s^j\right); j = 1,2,3,... \tag{11}$$

where, j portrays the layer index, A_g^j indicates the feature map index, A^{j-1} and X^j indicates the input as well as current layer's mask, E_{gr} specifies the trainable kernels. The 'Shepard interpolation' layers are weighted together to represent an interpolation operative of extremely nonlinear function. Here, s portrays the bias, ω specifies the network non linearity, and A indicates the smooth and differential function. Thus, the predicted outcome of 'ShCNN' is indicated as J_x^* in which the weight of 'ShCNN' is given training utilizing IIHBOA.

3.4 Recommended IIHBOA for training the 'ShCNN'

This portion defines the training procedure of 'ShCNN' using IIHBOA, which is crafted by the consolidation of IIWO [10] and HBO [11] technique. HBO [11] is an optimization approach, which is devised by mimicking the honey searching and digging characteristics of honey bees. The advantage of HBO algorithm is that the exploration rate of this scheme was high. Likewise, IIWO

is an extension of IWO, which is designed by utilizing the behaviour of weed colonies. The advantage of IIWO algorithm is that the security of IIWO is high. Thus, the devised IIHBOA scheme is developed by considering the benefits of both IIWO and HBO technique. Furthermore, the technical steps of devised IIHBOA are elaborated as below.

i) Initialization

In the initial step, the honey badger, and its locations are initialized as below.

$$m_i = J_i + n_1 \times \left(W_i - J_i\right) \tag{12}$$

where, m_i indicates the position of honey badger, W_i shows the higher limit and J_i provides the lower limit.

ii) Fitness factor determination

The fitness factor determination is evaluated to ascertain the optimised solution. In this investigation, the fitness function with smallest possible estimation of MSE is selected an as an optimal possible answer, which is stated by,

$$Fitness = \frac{1}{\varpi}\sum_{x=1}^{\varpi}\left[J_x - J_x^*\right] \tag{13}$$

where, J_x describes the expected outcome, J_x^* shows the predicted outcome, and ϖ specifies the total sample count.

iii) Description of intensity

Here, the term intensity depends on the hardiness of prey and the separation across the prey as well as badger, which is expressed as,

$$Q_i = n_2 \times \frac{I}{4\pi d_i^2} \tag{14}$$

$$I = \left(m_i - m_{i+1}\right)^2 \tag{15}$$

$$d_i = m_{prey} - m_i \tag{16}$$

where, n_2 depicts the random number among 0 and 1, I indicates the strength of source, Q_i indicates the prey's smell intensity, and d_i describes the distance amongst the prey as well as badger.

iv) Renew the component of density

The component of density is utilized to assure the conversion among exploitation as well as exploration. The minimum count of iterations decreases the density factor κ, which is portrayed as,

$$\kappa = S \times \exp\left(\frac{-u}{u_{\max}}\right) \tag{17}$$

where, S be the constant.

v) Get away from the local optimal

This step is utilized to escape from the areas of local optima. Moreover, this algorithm considers the flag B , which modifies the search direction for providing the high chances tosearch the search-space rigorously.

vi) Update the location of agents

The position of agents is updated based on two stages, such as digging stage as well as honey stage.

a) *Digging stage*: In the digging stage, the badger honey operates like in Cardioid structure, and the Cardioids motion is stated as,

$$m_{new} = m_{prey} + B \times \sigma \times V \times m_{prey} + B \times n_3 \times \mu \times d_i \times \left|\cos(2\pi n_4) \times \left[1 - \cos(2\pi n_5)\right]\right| \quad (18)$$

where, m_{prey} depicts the location of prey, n_3, n_4 and n_5 indicates the random numbers, μ indicates the time changing search influence factor and σ represents the ability of honey badger.

where,

$$di = m_{prey} - m_i \quad (19)$$

Moreover, the flag B is given by,

$$B = \begin{cases} 1 & ,if\ n_6 \le 0.5 \\ -1 & ,else \end{cases} \quad (20)$$

where, n_6 denotes the random number among 0 and 1, and B depicts the flag, which is used to determine the optimal location of prey. Substitute formula (19) in formula (18), then the expression becomes,

$$m_{new} = m_{prey} + B \times \sigma \times V \times m_{prey} + B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right|(m_{prey} - m_i) \quad (21)$$

$$m_{new} = m_{prey} + B \times \sigma \times V \times m_{prey} + B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right| m_{prey} - B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right| m_i \quad (22)$$

$$m_{new} = m_{prey}\left[1 + B \times \sigma \times V \times + B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right|\right] - B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right| m_i \quad (23)$$

From IIWO [10], the optimal weeds are utilized to shift remaining weeds to the optimal location, which is given by,

$$m_i^{b+1} = \gamma(b) * m_i^b + \left(m_{best} - m_i^b\right) \quad (24)$$

Let us assume,

$$m_i^{b+1} = m_{new} \quad (25)$$

$$m_i^b = m_i \quad (26)$$

Apply equation (25) and equation (26) in equation (24), then the equation can be written as,

$$m_{new} = \gamma(b) * m_i + \left(m_{best} - m_i\right) \quad (27)$$

$$m_{new} = m_i\left(\gamma(b) - 1\right) + m_{best} \quad (28)$$

$$m_i = \frac{m_{new} - m_{best}}{\gamma(b) - 1} \tag{29}$$

Moreover, the digging stage is updated by applying formula (29) in formula (23), then the expression becomes,

$$m_{new} = m_{prey}\left[1 + B \times \sigma \times V \times + B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right|\right] - B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right|\left(\frac{m_{new} - m_{best}}{\gamma(b) - 1}\right) \tag{30}$$

$$m_{new} = \frac{(\gamma(b) - 1) m_{prey}\left[1 + B \times \sigma \times V \times + B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right|\right] - B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right| m_{best}}{\gamma(b) - 1 + B \times n_3 \times \mu \times \left|Cos(2\pi n_4) * (1 - Cos(2\pi n_5))\right|} \tag{31}$$

This is the final updated equation of digging stage.

b) Honey stage:

The badger honey obeys the bird honey guide in order to arrives the beehive, which is expressed as,

$$m_{new} = m_{prey} + B \times n_7 \times \mu \times di \tag{32}$$

where, n_7 denotes the random number among 0 and 1,

$$di = m_{prey} - m_i \tag{33}$$

To boost functionality of honey badger algorithm, the honey stage is renewed using IIWO. Let us apply equation (31) in equation (30), then the equation is,

$$m_{new} = m_{prey} + B \times n_7 \times \mu \times \left(m_{prey} - m_i\right) \tag{34}$$

$$m_{new} = m_{prey}\left(1 + B \times n_7 \times \mu\right) - B \times n_7 \times \mu \times m_i \tag{35}$$

Substitute formula (29) in formula (35), then the expression is given by,

$$m_{new} = m_{prey}\left(1 + B \times n_7 \times \mu\right) - B \times n_7 \times \mu \times \left(\frac{m_{new} - m_{best}}{\gamma(b) - 1}\right) \tag{36}$$

$$m_{new} = \frac{m_{prey}\left(1 + B \times n_7 \times \mu\right)\left(\gamma(b) - 1\right) + B \times n_7 \times \mu \times m_{best}}{\gamma(b) - 1 + B \times n_7 \times \mu} \tag{37}$$

This is the final updated equation of honey stage process. where, m_{prey} signifies the prey lion of honey badger.

$$\mu = \Re \times \exp\left(\frac{-b}{\max}\right) \tag{38}$$

where, $\Re$ denotes the constant, which is greater than 1.

$$\gamma(b) = \left(\frac{T - b}{T}\right)^d \left(\gamma_{initial} - \gamma_{final}\right) + \gamma_{final} \times z(b) \tag{39}$$

where, $z(b)$ indicates the chaotic mapping.

$$B = \begin{cases} 1 & if\ n_6 \le 0.5 \\ -1 & else \end{cases} \tag{40}$$

vii) Reliability computation

The finest possible answer is gotten with fitness, and signified in formula (13), and possible answer with minimum fitness function is characterized as best possible answer.

viii) Termination

The aforementioned stages are performed uninterruptedly till uppermost solution is attained. Table 1 demonstrates the pseudo-code of IIHBOA.

Table 1. IIHBOA Pseudo-code

1	**Input:** Entire population m
2	**Output:** Best solution m_{new}
3	Initiate population
4	Compute fitness using equation (13)
5	**while** $u \le u_{\max}$ do
6	Renew the lessening factor κ using (17)
7	**for** $i = 1 to\ Y$ do
8	Measure intensity Q_i using (14)
9	**if** $n < 0.5$ then
10	Renew the location with equation (33)
11	**else**
12	Renew the location with equation (41)
13	**end if**
14	Compute new location and allocate to m_{new}
15	**if** $t_{new} \le t_i$ then
16	Assume $m_i = m_{new}$ and $t_i = t_{new}$
17	**end if**
18	**if** $t_{new} \le t_{prey}$ then
19	Set $m_{prey} = m_{new}$ and $m_{prey} = m_{new}$
20	**end if**

21 **end for**

22 **end while**

23 Evaluate the solution feasibility

24 Get m_{best}

Thus, the developed algorithm is used to train the weights of IIHBOA, which is developed by integrating the advantage of IIWOA and HBA. Moreover, the developed technique is used to enhance the selection of optimal region for embedding.

3.5 Wavelet transform

This section explains the process of wavelet transform, which is utilized to partition the input photo into eight sub layers using double decomposition. Here, the input of wavelet transform (WT) is the extracted key frames, which is partitioned into eight sub bands using wavelet transform. Moreover, the process of wavelet transform is explained as below.

WT: In this research, WT is applied to partition the video frames into 16 sub layers using two stage degradations. In the first level decomposition, the extracted key frame is initially partitioned into four sub layers, such that LL, HL, HH, and LH are generated. Here, LL characterizes coarse stage coefficients, LH, HH, and HH specifies the highest degree wavelet coefficient. Therefore, the LL sub layer is utilized to attain the second layer of decomposition. To progress the image excellence, the secret audio is embedded with LH, HH, HL, and LL sub layers. Thus, sub layers are articulated as,

$$\{LL, LH, HL, HH\} = S(P_i) \tag{41}$$

where, P_i specify the key frame, S designate the wavelet transformation, and $\{LL, LH, HL, HH\}$ are the decomposed wavelet coefficients. Each layer of photo P_i is processed under second layer of degradation for creating 16 sub-layers and are articulated as,

$$\{(LL)_4, (LH)_4, (HL)_4, (HH)_4\} = S(HH) \tag{42}$$

$$\{(LL)_1, (LH)_1, (HL)_1, (HH)_1\} = S(LL) \tag{43}$$

$$\{(LL)_3, (LH)_3, (HL)_3, (HH)_3\} = S(HL) \tag{44}$$

$$\{(LL)_2, (LH)_2, (HL)_2, (HH)_2\} = S(LH) \tag{45}$$

Thus, LL and HH coefficients are utilized to initiate embedding as well as extraction process. Moreover, the partitioned key frames are characterized as S_o.

3.6 Embedding secret audio message into object location

It is the process of inserting secret audio message into the selected region. The input of embedding process is the secret audio message, wavelet transform output and the predicted region

of object. Here, the audio message undergoes process of binary conversion using QRDA. Since, the embedding process is performed only on the binary values. Thus, the audio signal is needed to be converted into binary format in order to perform the embedding using QRDA.

3.6.1 QRDA

The QRDA [14] scheme is used to transform the audio signal into binary data. Generally, the audio signal contains both the amplitude and time information in analog form, which is hard to process. Hence, the amplitude as well as time information is transformed into the binary or digital format using QRDA, which is easy to execute. An audio signal is assumed as,

$$M = m_0, m_1, ..., m_{Y-1} \tag{46}$$

where, Y depicts the audio length. QRDA utilizes the q-bit sequence, which is given by,

$$H_P = H_P^0 H_P^1 ... H_P^{h-2} H_P^{h-1} \tag{47}$$

The encoded form of amplitude value is given by,

$$c_p = H_P^0 H_P^1 ... H_P^{h-2} H_P^{h-1}; H_P^e \in \{0,1\}; e = 0,1,...,h-1 \tag{48}$$

where, $P = 0,1,..., Y-1$ indicates the information of time. Hence, the audio signal can be rewritten as,

$$\left|C\right\rangle = \frac{1}{\sqrt{2}^y} \sum_{P=0}^{Y-1} \left|H_P\right\rangle \otimes \left|P\right\rangle \tag{49}$$

$$\left|P\right\rangle = \left|l_0 l_1 ... l_{y-1}\right\rangle, l_e \in \{0,1\} \tag{50}$$

$$\left|H_P\right\rangle = \left|H_P^0 H_P^1 ... H_P^{h-2} H_P^{h-1}\right\rangle; H_P^e \in \{0,1\} \tag{51}$$

where,

$$y = \begin{cases} \lceil \log_2 Y \rceil & Y > 1 \\ 1, & Y = 1 \end{cases} \tag{52}$$

where, $\otimes$ depicts the notation of tensor product, $\left|H_P\right\rangle = \left|H_P^0 H_P^1 ... H_P^{h-2} H_P^{h-1}\right\rangle$ indicates the binary consequence of amplitude and $\left|P\right\rangle = \left|l_0 l_1 ... l_{y-1}\right\rangle$ indicates the time information of audio. Thus, the QRDA requires $h+1$ bits to indicate the audio message. The binary audio signal is illustrated as A_m.

After the binary data conversion, the converted binary data is embedded into the selected region. Here, the embedding is done by combining both the secret message with the decomposed sub bands and predicted region of object.

$$E_f = S_o + \beta \times A_m \tag{53}$$

where, E_f depicts the embedded message, S_o indicates the partitioned key frames, β depicts the strength of embedding and A_m indicates the binarized audio message from QRDA. Then, an

inverse wavelet transform is subjected into the extracted frames in order to attain the embedded video, which is indicated as E_f .

Phase II. Extraction phase

In the extraction stage, the embedded multimedia is considered as an input. From the embedded video, the video frame is extracted, and then the object extraction and region prediction is carried out using active contour and IIHBO-based 'ShCNN'. On the other hand, the wavelet transformation is applied on the extracted object. Finally, the extraction process is done by utilizing the wavelet transform and predicted region so as to retrieve the binary audio. Moreover, QRDA is applied on the binary audio such that the secret audio message is extracted. Figure 2 shows the extraction phase of devised scheme.

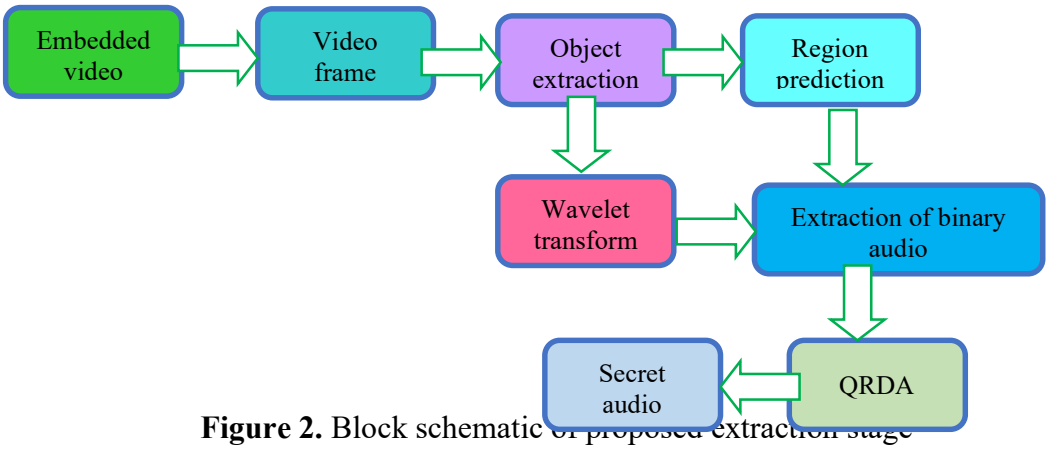


Figure 2. Block schematic of proposed extraction stage

To retrieve the covert communication from embedded video, the following steps are performed.

- Video key frame extraction
- Object extraction
- Region prediction
- Wavelet transformation
- Audio message transformation
- QRDA retrieving

All the above processes are performed as the same manner in embedding phase. The only difference is that, the extraction process considers embedded video as the input but the embedding process considers original video as input. At last, the QRDA retrieving is performed to regain the original audio message, which is explained as below.

QRDA retrieving: It is the process of retrieving the embedded audio information [14]. In this method, it recovers all the samples separately, which is expressed as,

$$Q = \sum_{q=0}^{2h-1} q|q\rangle\langle q \tag{54}$$

It is utilized to retrieve the values of quantum state amplitude of audio. Thus, the recovered audio samples are indicated as,

$$\langle H_P|Q|H_P\rangle = \langle H_P|\left(\sum_{q=0}^{2h-1} q|q\rangle\langle q\right)|H_P\rangle \tag{55}$$

$$\langle H_P|Q|H_P\rangle = \sum_{q=0}^{2h-1} q\langle H_P\|q\rangle\langle q\|H_P\rangle = H_P \tag{56}$$

where, P indicates the amplitude of sample. In this method, it recovers every sample precisely, which indicates that the whole audio is to be retrieved.

4. Evaluation and findings

The description of the findings of IIHBO-based 'ShCNN' scheme is described in this portion. Furthermore, the experimental arrangement, representation of a dataset, comparative methods and their performances are included in this section.

4.1 A test-run arrangement

The IIHBO-based 'ShCNN' strategy is incorporated in MATLAB R2020a tool using PC with Windows 10 OS Intel i3 core processor.

4.2 Representation of a Dataset

The IIHBO-based 'ShCNN' method utilizes two datasets, namely Video Watermarking dataset [12] and audio dataset [13].The video watermarking dataset contains multiple numbers of frames, which is very robust against the harmful attacks. Moreover, this dataset contains the multiple numbers of frames acquired from video. ELSDSR is an audio dataset, which contains the audio signals of 3 speakers, 4 speakers and 5 speakers.

4.3 Measurement Statistics

The measurement statistics used for the experimentation of IIHBO-based 'ShCNN' scheme are CC, PSNR and MSE.

CC: It is a metric, which is utilized to calculate the relationship among two variables, and is portrayed as,

$$CC = \frac{\sum\left(e_x - \overline{e}\right)\left(f_x - \overline{f}\right)}{\sqrt{\sum\left(e_x - \overline{e}\right)^2 \sum\left(f_x - \overline{f}\right)^2}} \tag{57}$$

where, CC signifies the correlation coefficient, e_x indicates the estimations of e-variable, $\overline{e_x}$ indicates the mean estimations of e-variable, f_x demonstrates the estimations of y-differential and $\overline{y_x}$ signifies the mean estimations of y-differential.

PSNR: PSNR is a metric, which is determined by taking the ratio among highest probable 'power of signal' to the 'power of corrupting noise', which is expressed as,

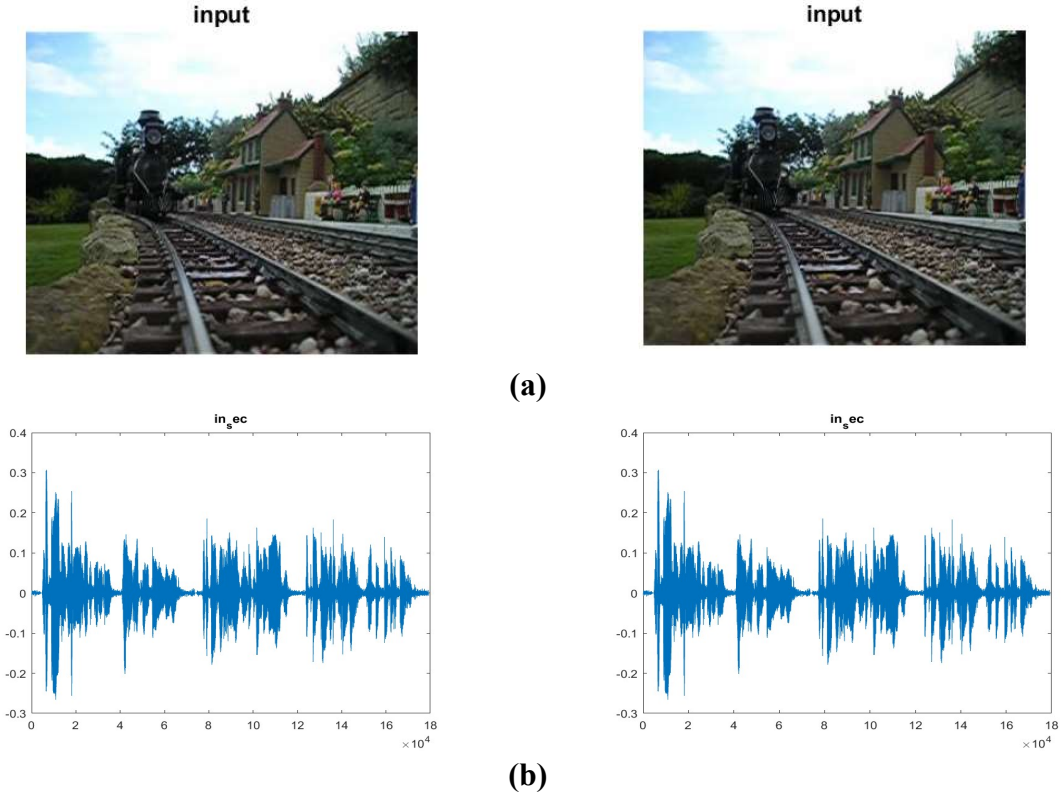
$$P = 20\log_{10}\left(\frac{Max_p}{\sqrt{E}}\right) \tag{58}$$

where, P signifies the PSNR, Max_p indicates the maximum power, and E signifies the MSE values of signal.

MSE: MSE is a metric, which is expressed as the mean of square values of error, which is already explained in equation (13).

4.4 Investigational Findings

The investigational findings of IIHBO-based 'ShCNN' scheme are shown in figure 3. The input frame acquired from the video is given in figure 3a) and the speech signal utilized for performing the immerging procedure is given in figure 3b). After the completion of embedding, the embedded image is to be obtained, which is demonstrated in figure 3c) and the original speech signal obtained from the embedded picture is given in figure 3d).



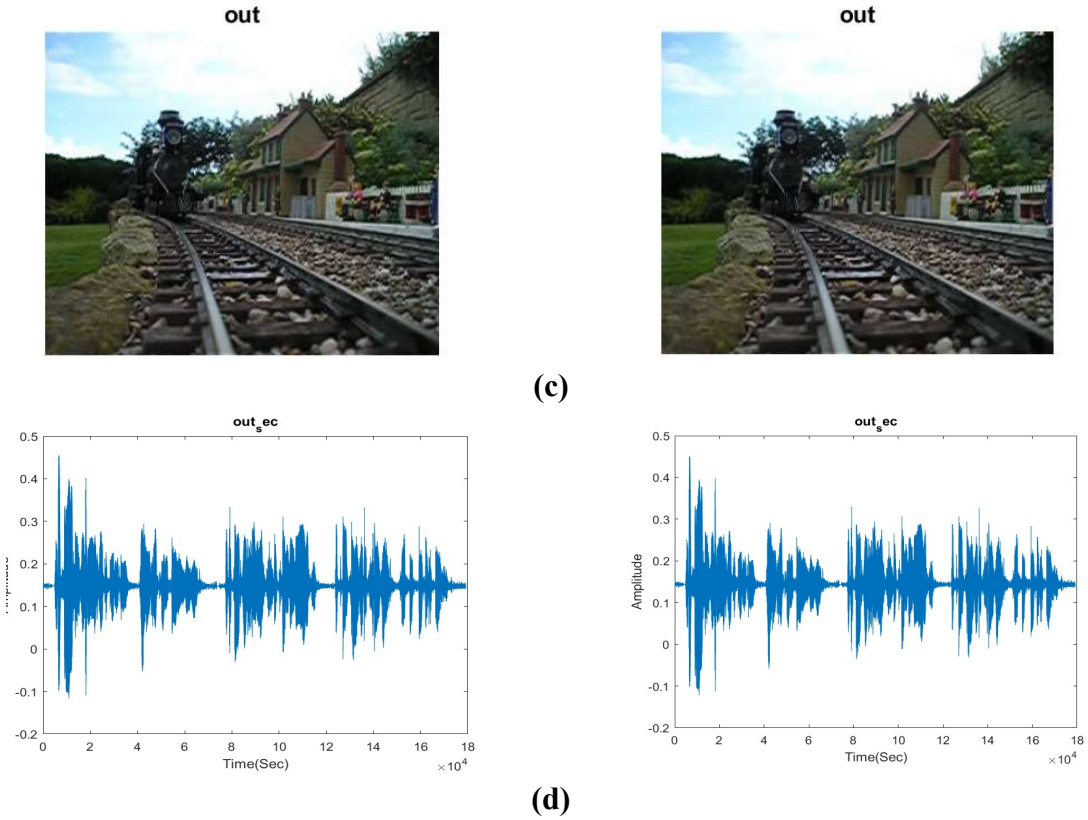


Figure 3. Investigational Findings a) Input frames, b) Input speech signal to embed, c) Embedded image, d) Extracted speech signal from embedded image

4.5 Comparative methods

The comparison of devised IIHBO-based 'ShCNN' method is done by comparing it with some conventional approaches, such as Botnet model [1], DWT [3], Chaos-based video steganography [5] and DST-SBPNRM [6].

4.6 Comparative assessment

In this paper, the correlating assessment of devised IIHBO-based 'ShCNN'technique is done with random noise and noise free analysis.

4.6.1 Assessment with random noise

The efficacy of devised IIHBO-based 'ShCNN' is assessed by utilizing the dissimilar frames with random noise based on evaluation parameters, which is portrayed in figure 4. The analysis of projected IIHBO-based 'ShCNN' scheme in terms of CC is demonstrated by Figure 4a). Here, the CC of IIHBO-based 'ShCNN' is 0.816 when the frame count is 1, whereas the CC of Botnet model is 0.680, DWT is 0.722, Chaos-based video steganography is 0.780 and DST-SBPNRM is 0.781 for the frame count is 1. Figure 4 b) illustrates the analysis of MSE with existing and projected schemes. For the count of frame is 4, then the devised IIHBO-based 'ShCNN' achieved the MSE of 0.301, while the existing techniques acquired the MSE of 0.852, 1.564, 1.047

and 1.093. The PSNR of devised IIHBO-based 'ShCNN' is given in figure 4 c). The IIHBO-based 'ShCNN' with random noise acquired the PSNR of 36.58 and the conventional techniques with random noise achieved the PSNR of 28.76 dB, 25.67 dB, 31.27 dB and 33.68 dB for the number of frame is 5.

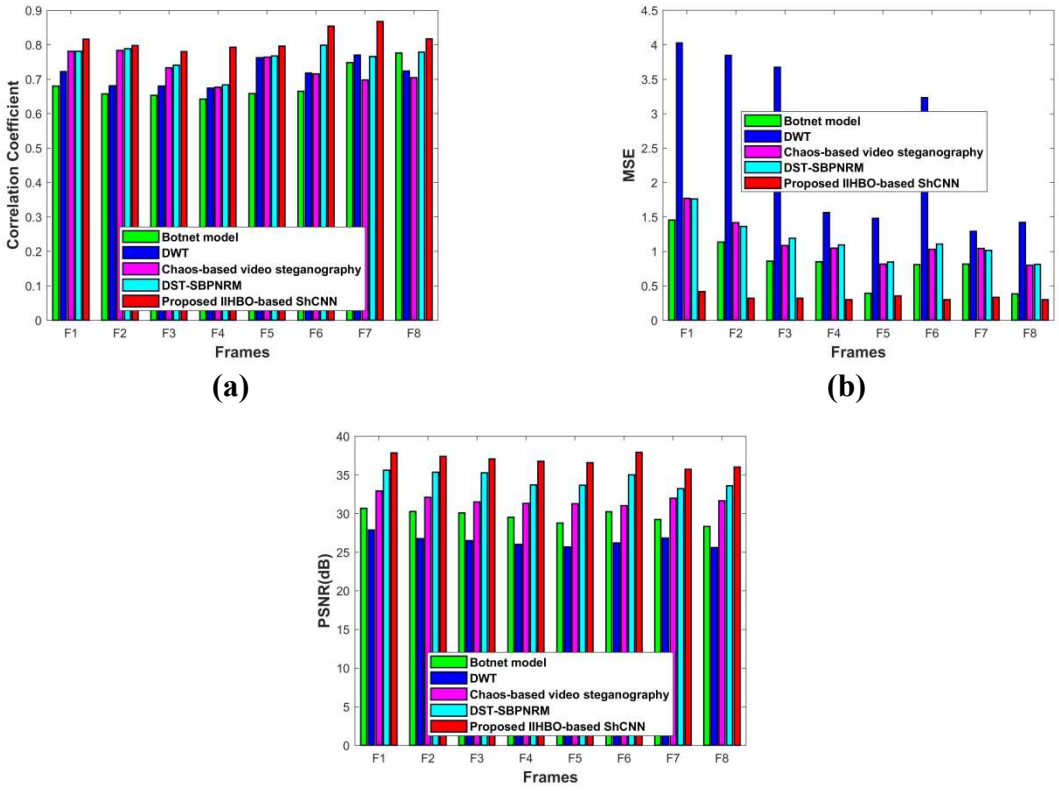


Figure 4. Analysis of devised IIHBO-based 'ShCNN' method with random noise in terms of a) CC, b) MSE, c) PSNR

4.6.2 Assessment based on without noise

The comparison of devised IIHBO-based 'ShCNN' technique with conventional techniques without the existence of noise is given in figure 5. Figure 5a) illustrates the assessment of CC with existing and projected schemes. For the count of frame is 4, then the devised IIHBO-based 'ShCNN' achieved the CC of 0.913, while the existing techniques acquired the CC of 0.747, 0.754, 0.810 and 0.848. The MSE of devised IIHBO-based 'ShCNN' is provided in figure 5 b). Figure 5 b) displays the assessment of projected IIHBO-based 'ShCNN' scheme in terms of MSE. Here, the MSE of IIHBO-based 'ShCNN' is 0.029 when the frame count is 1, whereas the MSE of Botnet model is 0.066, DWT is 0.142, Chaos-based video steganography is 0.089 and DST-SBPNRM is 0.086 for the frame count is 1. The PSNR of devised IIHBO-based 'ShCNN' is given in figure 5 c). The IIHBO-based 'ShCNN' acquired the PSNR of 43.51 dB and the conventional techniques with random noise attained the PSNR of 35.47 dB, 33.29 dB, 38.28 dB and 40.09 dB for the number of frame is 5.

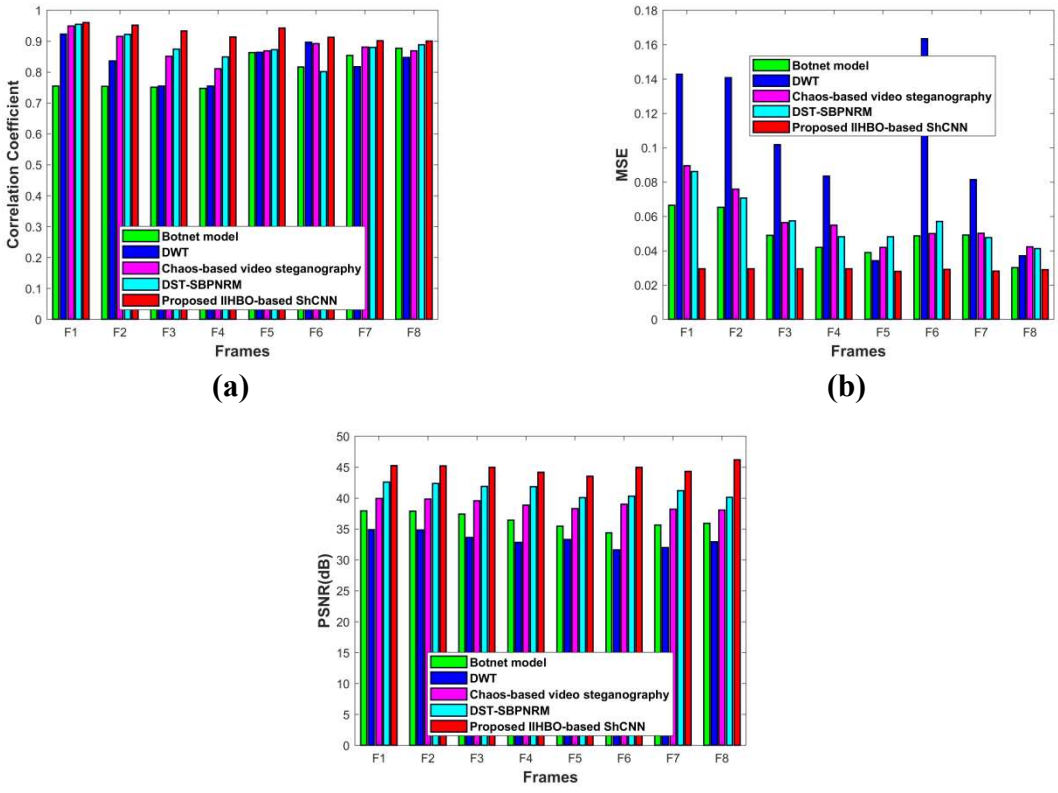


Figure 5. Evaluation of proposed IIHBO-based 'ShCNN' method with the absence of noise in terms of a) CC, b) MSE, c) PSNR

4.7 Correlating assessment

Table 1 shows the correlating assessment of devised IIHBO-based 'ShCNN' scheme. In this research, the analysis is done by changing the number of frames with the existence and inexistence of noise. From the table, it is plainly demonstrated that the IIHBO-based 'ShCNN' approach attained the better with the absence of noise in terms of CC, MSE and PSNR of 0.900, 0.028 and 46.19 dB, correspondingly. Likewise, with the existence of random noise, the developed method achieved as CC, MSE and PSNR of 0.817, 0.302 and 36.00 dB.

In the developed IIHBO-based 'ShCNN' method, the better performance is attained without the existence of noise. From the table 1, as existing procedures are effective, the devised schemes perform better than conventional techniques. To be able to boost performance of embedding, the wavelet transformation is applied on the extracted frame for embedding. Moreover, the optimized deep learning approach is utilized to predict the object region in order to embed the audio signal in the predicted region, which improves the effectiveness of deep learning based on CC, MSE and PSNR.

Table 1. Comparative discussion

Variations	Evaluation metrics	Botnet	DWT	Chaos	DST-SBPNRM	Proposed IIHBO-based 'ShCNN'
Random noise	CC	0.776	0.724	0.704	0.778	0.817
	MSE	0.382	1.421	0.801	0.814	0.302
	PSNR	28.321	25.60	31.65	33.60	36.00
Without noise	CC	0.876	0.847	0.868	0.888	0.900
	MSE	0.0301	0.037	0.042	0.041	0.028
	PSNR	35.94	32.92	38.09	40.13	46.19

5. Conclusion

This paper formulates the deep learning based optimization, namely IIHBO-based 'ShCNN' for video object watermarking. In this research, the effectiveness of embedding is enhanced by predicting the object region. The prediction of object region is done with 'ShCNN', which is trained by the hybridized IIHBO technique. IIHBO is a hybrid optimization scheme, which is modelled by the incorporation of IIWOA and HBO. Moreover, the devised method embeds the binary form of audio signal on the video frame. Prior to embedding, the wavelet transform is applied on key frames, which is combined with the predicted region of object; thereby the embedded video is to be attained. In addition, the embedded audio signal from embedded video is extracted by applying the inverse wavelet transform. Especially, QRDA is applied on the secret audio message for transforming the audio message into binary form of audio message in order to achieve the effective embedding. Furthermore, with the existence of noise, the IIHBO-based 'ShCNN' method obtained the superior performance utilizing random noise in terms of the CC is 0.817, MSE is 0.302 and PSNR is 36.00dB. With the absence of noise, the IIHBO-based 'ShCNN' approach attained the CC, MSE and PSNR of 0.900, 0.028 and 46.19 dB, correspondingly. Going forward, the performance of IIHBO-based 'ShCNN' is moreover enhanced by adding some more efficient optimization methods.

References

[1] Kwak M, Cho Y., "A novel video steganography-based botnet communication model in telegram sns messenger", Symmetry, vol.13, no.1, pp.84, January 2021.
[2] Yao Y, Yu N., "Motion vector modification distortion analysis-based payload allocation for video steganography", Journal of Visual Communication and Image Representation, vol.74, pp.102986, January 2021.
[3] Dalal M, Juneja M., "A secure video steganography scheme using DWT based on object tracking", Information Security Journal: A Global Perspective, pp.1-8, March 2021.
[4] Zhao H, Liu Y, Wang Y, Liu S, Feng C., "A video steganography method based on transform block decision for H. 265/HEVC", IEEE Access, vol.9, pp.55506-21, February 2021.
[5] Siddiqui TJ, Khare A., "Chaos-Based Video Steganography Method in Discrete Cosine Transform Domain", International Journal of Image and Graphics, vol.21, no.02, pp.2150015, April 2021.

[6] Niu K, Li J, Yang X, Zhang S, Wang B., "Hybrid adaptive video steganography scheme under game model", IEEE access, vol.7, pp.61523-33, March 2019.
[7] MstafaRJ, Elleithy KM, Abdelfattah E., "A robust and secure video steganography method in DWT-DCT domains based on multiple object tracking and ECC", IEEE access, vol.5, pp.5354-65, April 2017.
[8] Patel R, Lad K, Patel M, Desai M., "A hybrid DST-SBPNRM approach for compressed video steganography", Multimedia Systems, vol.27, no.3, pp.417-28, June 2021.
[9] Ren, J.S., Xu, L., Yan, Q. and Sun, W., "Shepard convolutional neural networks", Advances in Neural Information Processing Systems, vol.28, pp.901-909, 2015.
[10] Misaghi M, Yaghoobi M., "Improved invasive weed optimization algorithm (IWO) based on chaos theory for optimal design of PID controller", Journal of Computational Design and Engineering, vol.6, no.3, pp.284-95, July 2019.
[11] Hashim FA, Houssein EH, Hussain K, Mabrouk MS, Al-Atabany W., ", Honey Badger Algorithm: New metaheuristic algorithm for solving optimization problems", Mathematics and Computers in Simulation, vol.192, pp.84-110, February 2022.
[12] Video Watermarking using DWT, https://in.mathworks.com/matlabcentral/fileexchange/52225-video-watermarking-using-dwt?s_tid=prof_contriblnk, accessed on February 2021.
[13] ELSDSR Audio dataset from "http://cogsys.compute.dtu.dk/soundshare/elsdsr.zip, accessed on February 2021.
[14] Wang, J., "QRDA: quantum representation of digital audio", International Journal of Theoretical Physics, vol.55, no.3, pp.1622-1641, 2016.
[15] R. J. Mstafa and K. M. Elleithy, "Compressed and raw video steganography techniques: a comprehensive survey and analysis", Multimedia Tools and applications, vol.76, pp.21749–21786, 2017.
[16] Min Wu, Bede Li, "Multimedia Data Hiding, Springer", 1st edition, 2003.
[17] Dasgupta K, Mondal JK, Dutta P. "Optimized video steganography using genetic algorithm (GA)", Procedia Technology, vol.10, pp.131-7, January 2013.
[18] Khongsit, R. and Rangababu, P., "Scalable discrete Tchebichef Transform for image/video compression", In proceedings of 2017 International Conference on Innovations in Electronics, Signal Processing and Communication (IESC), pp. 127-131, 2017.
[19] Lamba, S. and Nain, N., "Detecting anomalous crowd scenes by oriented Tracklets' approach in active contour region", Multimedia Tools and Applications, vol.78, no.22, pp.31101-31120, 2019.