SPATIAL TRANSFORMER NETWORKS FOR ADAPTIVE VIDEO ENCODING AND DECODING IN VIDEO TRANSMISSION

AJITHA G1, SANTHIPRABHA I2, AJITHA G3

1 Research Scholar, Dept. of Electronics and Communications, University College of Engineering, JNTUK, Kakinada, India
2 Professor, Dept. of Electronics and Communications, University College of Engineering, JNTUK, Kakinada, India
3 Assistant Professor, Department of ECE, Institute of Aeronautical Engineering, Hyderabad

Email: ajithagphd@gmail.com

ABSTRACT

Video encoding is one of the methods for improving video transmission quality. Recently adaptive and scalable video encoding technology is emerging which compresses the video files while maintaining their quality. However, existing methods of video encoding and decoding are used to extract the contextual information, which does not suit for all kinds of devices, applications and services in wireless networks. Hence, the new video encoding and decoding technique is presented for video quality improvement using deep learning. For that purpose, Spatial Transformer Networks (STN) is presented in this paper. Firstly, coefficients are extracted in the set of frames of a given input video. This is done by means of fast curvelet transform method that splits the frame into multiple levels. Secondly, approximate coefficients generation is implemented by means of spatial transformer networks. Then the detailed coefficients are optimized by means of the Human Mental Search Optimization technique. Later, the side information is extracted using Inverse Curvelt Transform. Finally, the video quality is evaluated using Fuzzy Logic Inference System (FLIS) using video quality assessment metrics. This feedback is forwarded to the source node. The comparison is between the proposed approach with the existing approaches in terms of power consumption, time, PSNR, MSE and RMSE for different datasets.

Keywords: Video Encoding And Decoding, Spatial Transformer Networks, Fuzzy Logic, Side Information Extraction, Video Quality Assessment

1. INTRODUCTION

Multiple video applications and services, such as video telephony, HD and Ultrahigh-Definition (UHD) broadcasting, Internet protocol television (IPTV), and emerging mobile streaming, have advanced rapidly in the last decade. As a result, assessed the effectiveness of video that are broadcast and watched on youtube becomes an important research area. Video streaming has become extremely popular on the Internet, accounting for more than 55 percent of all traffic. Numerous scholars have put in a lot of effort to examine the efficiency of streamed communication systems. Similarly, organisations such as the International Telecommunications Union (ITU) have many frameworks and standardisation processes in place for evaluating observed image quality in a range of application scenarios. Numerous media will not accept video signals recorded with contemporary devices for transmission or storage. As a result, and during encoding stage, it is important to reduce its size. It is possible to achieve this by lowering spatially and/or temporal resolution. From the other hand, multimedia displaying technologies (mobile phones, tablet devices, laptop, etc.) are rapidly evolving and can exhibit high definition video signals regardless of screen resolution. As a result, in during decoding, it really is essential to boost the spatial and/or temporal resolution of the video file (prior to presentation). As a result, interest in spatial/temporal video upscaling techniques has exploded in the previous decade. First, frames from an uncompressed video feed are eliminated in this study. With both the increasing demands for images and multimedia applications, the need for uniform images or video standard evaluation metrics has grown. Different ways to estimating the detection
performance of images and videos have been offered in the research. DVC's encoding procedure is intentionally simple. DVC's encoding procedure is intentionally simple. First, the incoming video sequence is divided into groups based on the cumulative movements of all the images in each block over a predetermined threshold. A GOP is the sampling frequency in each of these groups (Group of Pictures). The first frame in a GOP is considered as the core frames, and the other frames are known as Wyner-Ziv (WZ) frames. The H.264 main profile intra coder will be used to extract the so-called key frames. A larger GOP decreases the data rate by increasing the number of WZ frames between crucial images. The so-called WZ frame will be transformed block by block, with DCT performed to each 4x4 block. DCT coefficient bands will be formed by grouping the DCT coefficients of the full WZ frame. Every coefficients bands will be homogeneous vector quantization with pre-defined levels that after transformation coding. On the quantization bins, bitplane ordering will be accomplished. The Low Density Parity Check Accumulator (LDPCA) encoding will be used to compress each ordered bit-plane independently. The cumulative disorder of the encoded bit planes is represented by a series of binary representation computed by the LDPCA encoder. To guarantee adequate decoding, an 8-bit Cyclic Redundancy Check (CRC) sum will be given to the decode for every image pixel. These partial products will be held in a cache in the encoder and sent out in stages to the decoding, which will continuously demand additional bits through into the feedback connection in decode. The DVC decode method seems to be more difficult. H.264 main profile intra decoder will be used to decoding the key frames. The decoded key frames will be used to recreate side information (SI) at the decoding, that is a guess at the WZ frame that is only accessible at the encoding. For SI generating, a movement adjusted approximation in between two nearest reference frames is conducted. Connection distortion in the virtual channel can cause a discrepancy in between WZ frame and the associated SI. To produce a good estimation of the remainder, an online Laplacian model is applied (WZ-SI). The SI will be subjected to the DCT transform, yielding an approximation of the WZ frame's coefficients.

![Fig.1. DVC based Encoder and Decoder Architecture](image)

Soft input data for the information bits will be produced to use these DCT coefficients, taking into consideration the mathematical techniques of the artificial noises. By taking into account the earlier encoded bit planes and the value of input data, the probabilistic model produced for every Frequency component will be translated into unconditional bit probabilities. The presented method's goals are as follows:

- Perform encoding, which is the process of compressing content so that it requires less time;
- Compressing moving frames through video processing is yet another goal
- Ensure the quality of the videos at the destination side

This paper's contributions can be summarized as follows:

- Wavelet coefficient approximations are utilized to construct training (input, target) pattern data using a three-level decomposition using fast curvelet transform on both even and odd pre-defined counts of frames.
• Spatial transformer network (STNs) for the utilization of approximate coefficients extraction for the input frames
• Further the frames are optimized by means of inverse curvelet transform. This provides the SI frame information from previously computed estimated and detailed coefficients for each sub-band.
• Finally, the fuzzy logic inference system is used for video quality estimation at the destination node. Depends upon the information of destination node, the source node improves the quality of video by changing of bit rate adaptively by means of human mental search optimization algorithm.

2. LITERATURE REVIEW

With simultaneous and true decode in DVC, estimate of rapid and consistent side information is a significant issue. For elevated videos, the problem gets considerably more acute. To decrease the total cost of side information prediction, a computationally simple DVC codec is suggested in this paper, which uses a basic phase interpolation (Phase-I) technique. It is fast for all acquired images, and it produces considerable results for high-resolution recordings with such a large number of images (GOP). With just an increase in precision, the computing time for the proposed methods dramatically lowers. For high-resolution films and large GOP, it is 221 percent to 280 percent faster than standard frame interpolation, with little reduction in the visual clarity of predicted side information [21]. Side information (SI) is the predicted frames of the actual Wyner-Ziv (WZ) frame inside DVC systems, and it has a huge effect on the overall system operation. The decoder uses high-quality SI to reconstruct Wyner-Ziv frames. As a result, we suggest a novel SI generation approach based on a study of existing SI generation systems. Motion compensation frame interpolation (MCFI) and optical flow approximation provide two kinds of SI, which would then be combined using a probabilistic fusing technique to gather highly precise SI. The experimental findings show that the Hybrid SI generating algorithm suggested in this paper may successfully increase SI quality [22]. In DVC, high-quality side information is essential for trustworthy soft-input information in the decoding to decode every DCT band of a Wyner-Ziv (WZ) frame, resulting in much more effective decoding. As a result, fewer error-correcting bits must be transferred from the encoding to the decoding to decoding each DCT band’s bitplanes, resulting in improved compression effectiveness and rate-distortion performance. In this research, researchers look at how to enhance the rate capability of distributed video coding by gradually enhance the effectiveness of side information frames. A novel paradigm for consecutive side information refinement is described, which uses the specific data received after decode the preceding DCT bands of a WZ frame to improve the greater initial information frame. The matching side information frames is improved when additional information about the WZ frame becomes known after each DCT band of the WZ frame is decoded, and then used to decoding the next DCT band of the WZ frame. Simulations show that the suggested approach for side information frame refining improves the RD efficiency of distributed video coding significantly [23]. As comparison to progressive video coding, the DVC system is shown to have no performance degradation in theory. Unfortunately, there is a significant difference between mathematically maximum efficiency and real execution. At the decoding, the side information (SI) - a noise version of the original Wyner-Ziv frame (WZF) - is created, and the correlations noise - the difference between both the actual WZF and matching SI - is modelled. If the SI & correlations noise are computed as correctly as feasible, the DVC method will perform more effectively. So, utilizing information within decode WZF’s as during decoding stage, this research presents a strategy to improve the quality of SI and also the correlations noise models. With each bitplane is decode, the original SI created by Movement Temporal Interpolation (MCTI) and previously decoded keyframes (KF) will be used as reference frames to improve the information sequentially. The experimental results showed that utilizing this strategy enhances the effectiveness of the distributed video coder considerably [24]. DVC is being looked at as a possible option for uplink applications including wireless video surveillance and multimodal sensor technologies. The accuracy of side information (SI) has a substantial impact on the codec's rate-distortion (RD) effectiveness in distributed video coding.
Nevertheless, the quality of the side information fluctuates throughout the sequence within each frame. As a result, in terms of improving side information, this research proposes a motion activity-adapted side information creation technique. The experiments demonstrate that utilising this strategy enhances the effectiveness of the distributed video coder substantially [25].

DVC enables adjustability in complex distribution seen between encoder and decoder. The decoder driven side information (SI) specifically for creating, like every building piece, is an important part of a DVC codec. The reliability of the SI produced at the decoder is crucial to the effectiveness of a DVC codec. The SI frame is a replica of the original Wyner-Ziv (WZ) frame. As a result, the higher the SI quality, the higher the codec's efficiency. The fundamental goal of this research is to increase the SI frame's quality in order to improve the DVC's current effectiveness. This paper mentions a hybrid SI generation technique that uses the principles of the discrete wavelet transform (DWT) and the extreme learning machine (ELM) algorithms in a transform domain-based DVC framework to achieve this goal.

3. PROPOSED WORK

In the proposed work, the SI method used in this study seeks to improve the reconstructed frame quality of videos. For video encoding and decoding, the proposed SI generation employs spatial transformer networks and a fast curvelet transform algorithm. Figure 1 depicts the flow of the suggested system model.
• **Video encoding**
  Initially source node encodes the video before transmitting it to destination. In video encoding, **fast discrete curvelet transform** scheme is employed. This process is performed in application layer since our work is based on cross-layer approach; application layer is able to communicate with physical layer and network layer. In our work, FCT can adjust the parameters (such as number of enhancement layers, and rate of group of pictures (GOP)) accordance to congestion and video quality.

• **Video quality estimation**
  Finally, the encoded video is received in the application layer of destination. Then the video is decoded and the quality of video is estimated by **Fuzzy Logic Inference System** scheme. In FLIS, peak to signal to noise ratio (PSNR), frame loss rate, and packet size difference are considered to estimate the quality of the video. After estimation, the **Video Quality Score (VQS)** is sent to source through **Real-time Transmission Control Protocol (RTCP)**.

A. **FDCT**

Then coefficients (coefficients) are extracted by FDCT. This transformation method is used to mine the coefficients in the image. In FDCT, frequency coefficients are extracted that purpose is to extract the coefficients. However, curvelet transform is denoted by multi-scale object representation. It implementation is applied over the resized image and choose the coefficients from that. With the use of FDCT, the input image is decomposed into 4 levels, and the coefficients are calculated. The curvelet coefficients are coefficient vectors, which considered improving the object determination in the image, which executes based on the Fourier samples wrapping. The curvelet transform consists of multi-scale pyramids with different directions and positions. The FDCT is defined as the following:

\[ c(j, l, k) = f, \Phi_\text{(j, l, k)} \]

(1)

Where \( j, l, k \) are defined as the parameters of scale, direction, and position. In FDCT, the image is given by the ROI segmented image, and the output will be the co-efficient extraction. Firstly, we initialize the data structure with shift condition. The input image is decomposed into the set of bands, in which Pyramid Scale Decomposition is determined using \( M_{\text{pyr1}} \) and \( M_{\text{pyr2}} \). The pyramid is utilized to provide the various sub-bands. Each sub-band is windowed smoothly into “squares”. Secondly, the smooth periodic extension of the high frequencies is mentioned in the Eqn.(4.1). Then we calculate the window scale and wrapping window to acquire the efficient curvelet coefficients. In wrapping, two processes are involved such as periodize and re-index the windowed frequency data. Thirdly, we calculate the normalization for \( x \)-coefficient, and \( y \)-coefficient by wrapping conditions. Later that, low-pass filter is applied. When the scale decomposition is completed, we determine the angular decomposition for horizontal, vertical...
bands, left and right bands. At last, we identify that the curvelet level by following function.

\[
\text{Curvelet}_{\text{wav}} = \frac{\text{FFT}_{\text{SHIFT}}(\text{IFFT} \ 2(l_{\text{wed}})) \ast \text{size}(roi_{\text{ex}})}{(2)}
\]

Algorithm 1: Fast Discrete Curvelet Transform

**Input:** Coefficient Extracted image (roi_{ex})

Let \( N_g \) be the Number of scales.

To initialize the data structure,

Finest, \( F_g = 3 \);

Let shift condition,

\( \text{img}_{\text{sh}} = \text{ifftshift}(\text{ifftshift}(roi_{\text{ex}})) / \sqrt{\text{size}(roi_{\text{ex}})} \)

\[ [N_e, N_v] = \text{size}(\text{img}_{\text{sh}}) \]

Where,

\( N_e \) be the Number of edges

\( N_v \) be the Number of vertex

To find the pyramidal scale decomposition

\( M_{\text{pyr1}} = \frac{N_e}{3}; \ M_{\text{pyr2}} = \frac{N_v}{3}; \)

To compute smooth periodic extension of high frequencies,

If \( F_g = 1 \)

---

**Fig. 3. Spatial Transformer Networks Flowchart**
SN_ε = 2 * floor(2 * M_{pyr1}) + 1
SN_η = 2 * floor(2 * \text{M}_{pyr2}) + 1

I_{gh} = I_{gh}(\text{equi}_{1d1}, \text{equi}_{1d2})

To compute windows length,
\[ w_{ten1} = \text{floor}(2 * M_{pyr}) - \text{floor}(M_{pyr1}) - 1 - \text{mod}(N_{e1}) = 0; \]
\[ w_{ten2} = \text{floor}(2 * M_{pyr2}) - \text{floor}(M_{pyr2}) - 1 - \text{mod}(N_{e2}) = 0; \]
\[ co_1 = 0; (1/w_{ten1}) : 1; \]
\[ co_2 = 0; (1/w_{ten2}) : 1; \]

To compute wrapping window,
\[ w_1 = \text{zeros(size(co_x))} \]
\[ w_1 = \text{zeros(size(co_x))} \]
\[ w_1(c_o_x \leq 0) = 1 \]
\[ w_1(c_o_x > 0) & (c_o_x < 1) = \exp(1 - \exp(1 / c_o_x))(c_o_x > 0) & (c_o_x < 1); \]
\[ w_1(c_o_x \geq 0) = 1 \]
\[ w_1(c_o_x > 0) & (c_o_x < 1) = \exp(1 - \exp(1 / c_o_x))(c_o_x > 0) & (c_o_x < 1); \]

Let find the low pass filter,
\[ L_{pf} = \left\lfloor \frac{w_1, \text{ones}(1, 2 * \text{floor}(M_{pyr1} + 1))} \right\rfloor \]

To compute angular decomposition,
\[ n_{qua} = 2 + 2^{\text{iter}}; \]
\[ \text{for } q_{re} = 1: n_{qua} \]
\[ M_{hor} = M_{pyr2} * (\text{mod}(q_{re}, 2) = 1) + M_{pyr1} * (\text{mod}(q_{re}, 2)) = 0; \]
\[ M_{ver} = M_{pyr1} * (\text{mod}(q_{re}, 2) = 1) + M_{pyr2} * (\text{mod}(q_{re}, 2)) = 0; \]
\[ \text{If } \text{mod}(n_{qua}, 2) \]
\[ W_1 = [W_{t, \text{left}}, W_{t, \text{right}}] \]
\[ \text{End} \]
\[ \text{End} \]

To compute left corner wedge,
\[ l = l + 1 \]

Let regular wedge,
To compute right corner wedge,
\[ \text{slow}_1 = \text{round}(2 * \text{floor}(4 * M_{ver}) * n_{qua} + 1) - \text{lcw}_1 / \text{floor}(4 * M_{ver}) \]
\[ \text{midc}_1 = \text{floor}(4 * M_{ver} + \text{floor}(M_{ver}) - \text{slow}_1 \]

Let regular wedge,
\[ l_{wed} = \text{floor}(4 * M_{ver} + \text{floor}(M_{ver}) + \text{slow}_1 \]

\[ W_{data} = \sqrt{2 * \text{img}(roi_{ex})} \]

To compute wavelet level,
\[ \text{curvelet}_{wav} = \text{fftshift} \left( \text{ifft2}(L_{wed}) \right) * \sqrt{\text{size}(roi_{ex})} \]
\[ \text{End} \]
\[ \text{End} \]

**Output:** Coefficient Extraction $F_{cur}$

---

**B. Spatial Transformer Networks**

After the coefficient extraction, the optimum set of coefficients are extracted by the spatial transformer networks, and also it removes the redundant coefficient vectors. There are three reasons to use the spatial transformer networks, which are following:

1. It considers the prior knowledge of the image (deep coefficients) for end-to-end training

2. It resolves the two issues of CNN: translation invariance and max-pooling layers

3. It considers only relevant coefficients and reduces the redundant coefficient vectors that significantly reduces the false positive rate

The spatial transformer networks address the translation invariance issues by performing two significant operations: Affine Transformation, and Interpolation. In traditional
CNN, this block is added to produces the effective outputs and finds the deep image coefficients for next step. It is well-suited for implementing for large and complex image cases. In addition to scale invariant issues, CNN have another important issues such as scale, background cluster and viewpoints variation. Let’s assume that the transformation matrix for the spatial transformer networks is \( AT_\theta \) and \( \theta \) is the linear transformation which is given as follows:

\[
\begin{bmatrix}
\theta_1 & \theta_2 & \theta_3 \\
\theta_4 & \theta_5 & \theta_6
\end{bmatrix} = \begin{bmatrix}
s_x & \varphi_x & t_x \\
\varphi_y & s_y & t_y
\end{bmatrix}
\]

\( (3) \)

Where \( s_x, s_y, \varphi_x, \varphi_y, \) are the skew parameters, \( t_x, t_y \) are the translation parameters. With the three sets of parameters, affine transformation matrix was calculated by the localization network. In this this grid generator computes the grid sampling\( G \in K^{kw\times w^2} \). It determines the arguments in the input data. This step finds the transformed output. We implements the grid warping on augmented input by sample kernel \( G_i = (x_i^s, y_i^s) \), which provides the coefficient map \( V \). Therefore the sample kernel which means bilinear sampling kernel is defined as follow:

\[
V^c = \sum_n \sum_{m} u^c_{n \times m} \max(0,1 - (x_i^s - m)\max(0,1 - |y_i^s - n|))
\]

\( (4) \)

Where \( x_i^s, y_i^s \) are the source coordinates of the augmented input \( U \) that represents the sample points. The final predicted result from the spatial transformer networks are transformed into CNN for final optimum coefficients set prediction

**C. Human Mental Search Optimization**

**Algorithm**

The clustering of extracted coefficients is performed in order to reduce the complexity in classification of modulation. For this purpose, a novel twin functioned-human mental search (TF-HMS) is proposed in this approach which was found to provide better performance than the Neutrosophic C-Means clustering algorithm. The HMS falls under population based optimization algorithm which utilizes bid space search. Initially, the populations of candidate solutions are generated randomly in the form of bid which are expressed as,

\[
Bid = \{x_1, x_2 \ldots x_d\}
\]

\( (5) \)

The evaluation of each bid is based on the specific objective function \( OF(bid) \) such as the quality of candidate solution which is represented as,

\[
OF(bid) = OF(x_1, x_2 \ldots x_d)
\]

\( (6) \)

The mental search operation is performed to group the bids in \( d \) dimensional space. The proposed TF-HMS considers the coefficient vectors \( \vec{b}_y \) as the form of bids which encode the cluster centre. The twin functions considered for clustering are cluster purity and cluster entropy function. The number of clusters is represented by array length which is denoted as \( K \). The upper and lower bound for each bid are computed which can be formulated as,

\[
L = \min(\vec{b}_y)
\]

\( (7) \)

\[
U = \max(\vec{b}_y)
\]

\( (8) \)

The formation of clusters is carried out based on three important coefficients namely inter-cluster distance, intra-cluster distance, and expression error respectively. The computation of weight of these parameters is represented as,

\[
OF(Bid, \vec{b}_y) = C_1d_{max}(Z, Bid) + C_2(Z_{max} - d_{min}(\vec{b}_y, Bid) + C_3d_E)
\]

\( (9) \)

Where, \( C_1, C_2, C_3 \) represents the weight values of parameters and \( Z_{max} \) denotes the maximum data value. The procedure of the proposed TF-HMS algorithm involves several stages which are described as follows,

**Stage 1:** Initially the parameters involved in the clustering process such as number of bids \( N \), number of clusters \( K \), number of clusters for grouping of bids \( K_{BG} \), and mental search parameters such as \( M_{min} \) and \( M_{max} \) are initialized.

**Stage 2:** Based on the number of bids \( N \), the population of the candidates \( P \) is formulated as,

\[
P = \begin{bmatrix}
Bid_1 \\
\vdots \\
Bid_N
\end{bmatrix}
\]

\( (10) \)

**Stage 3:** The objective function \( OF \) is computed and the evaluation of bids based on \( OF \) is performed

**Stage 4:** The selection of best bid is executed based on the evaluation result from previous stage.

**Stage 5:** The selection of random value is performed, in which a random value between \( M_{min} \) and \( M_{max} \) is selected for each bid.
Stage 6: The creation of new bids is carried out in the vicinity of prevailing bids by using levy function which can be formulated as,

\[ N_{POS} = Bid_i + S \]  

(11)

Where, \( S \) denotes the number of steps which is determined based on random values and maximum number of iterations (\( Max_i \)) respectively. The computation of \( S \) is formulated as follows,

\[ S = \left( 2 - I \times \left( \frac{2}{Max_i} \right) \right) \times \left( \lambda \otimes \text{Levy} \right) \]  

(12)

Where, \( I \) denotes the current iteration, \( \lambda \) denotes the random variables, and \( \otimes \) denotes the element-wise multiplication respectively. The step size of \( S \) can be derived as follows,

\[ S = \left( 2 - I \times \left( \frac{2}{Max_i} \right) \right) \times 0.01 \]  

(13)

\( U, V \) denotes two random variables, and \( x^* \) denotes the best position respectively.

Stage 7: The replacement of previous bid is carried out by the new bid when \( OF \) of new bid is better than the previous bid.

Stage 8: the grouping of bids into number of clusters is carried out in this stage based on the parameters.

Stage 9: In this step, the objective function of each cluster is computed and the evaluation of clusters is performed. The cluster possessing lower bound of mean objective function is considered as final cluster. The other clusters pass towards the best bid in the collection.

Stage 10: Go back to stage 2 and continue the other stages until the stop criterion is satisfied.

D. Fuzzy Logic Inference System

Fuzzy inference system is the key unit of a fuzzy logic approach, which is proposed for decision making as its main work. It uses the “IF...THEN” fuzzy rules to employ best decision making. The major characteristics of fuzzy inference system are following:

- The FIS output is always a fuzzy set irrespective of its input which can be crisp or fuzzy
- When it is considered as a controller, it must have defuzzified outputs
- In defuzzification, fuzzy variables are transformed into crisp variables.

Fundamental Elements of FIS are follows:
- **Database** – It defines the membership functions for fuzzy sets which is considered in rule base
- **Rule base** – It comprised of set of fuzzified rules (IF-THEN)
- **Decision Making Unit** – It allows to perform operations on rules
- **Fuzzification interface unit** – It transfers the crisp quantities into fuzzy quantities
- **Defuzzification interface unit** – It transfers fuzzified quantities in to crisp quantities

There are two methods are used in FIS that are: Mamdani Fuzzy Inference System and Takagi-Sugeno Fuzzy System (TS method). In Mamdani fuzzy inference system, output membership functions expect fuzzy sets. After the aggregation process, there is a fuzzy set for each output variable that requires defuzzification. The merits of Mamdani model are: (1). It is intuitive, (2). It has widespread acceptance and (3). It is fit for human input. In TS method, rule format is given as follows

\[ IF \ x \ is \ A \ and \ y \ is \ B \ THEN \ Z = f(x,y) \]  

(14)

where AB are fuzzy sets in antecedents and \( Z = f(x,y) \) is a crisp function in the consequent. Such type of IF-Then rule for a Sugeno model has the following form:

\[ IF \ input \ 1 = x \ and \ input \ 2 = y \ then \ Output \ is \ Z = ax + by + c \]  

(15)

For a zero order sugeno model, the output level \( Z \) is a constant value (\( a = b = 0 \)). The merits of Sugeno method are follows: (1). It can be well-suited in linear models, (2). It can be efficient and flexible for adaptive and optimization technique, (3). It ensures continuity of the output surface, and (4). It is computationally efficient. However, fuzzy inference system is a precise and effective problem-solving methodology. It is able to handle linguistic knowledge and numerical data.
In first FIS, we considered two input variables including bit rate and resolution. Totally there are $8 \left(2^3\right)$ fuzzy rules are used in first fuzzy inference system. In second FIS, two input variables are used including expected execution time and video frame rate. Here also we have 8 fuzzy rules. Finally, the sum of two FISs is added to compute the quality.

There are three linguistic states and term set of input variables for the double fuzzy inference system are shown low, high and medium. The five input variables are fuzzified using a trapezoidal fuzzifier that determines membership grade for each of the fuzzy sets. In inference module (the fuzzy rule base and inference), the fuzzified input measurements are then utilized by inference engine to measure control rules stored in the rule base based on the rule strength. Fuzzy
inference is the process of formulation of a mapping between two variables (input and output) using fuzzy logic algorithm. Various video quality enhancement parameters are used for the video decoding quality improvement.

4. EXPERIMENTAL EVALUATION

For evaluations, the proposed system was examined using MATLAB and conventional and accessible video sequences such as foreman and tennis sport. The input images were discovered to include a wide range of motion trajectories and speeds (from slow to fast), as well as multiple texture compositions, and were therefore used it to conduct a comprehensive evaluation of DVCs. Figure 5 shows the suggested spatial transformer networks-based SI estimation. The foreman video is used as an input in this presented design, including a prototype figure 6. Figure 7 shows the DCT frames that have been restored.

![Fig.6.GUI for side information estimation](image)

![Fig.7.Input (foreman)](image)
In terms of PSNR, MSE, and RMSE, the proposed FDCT with STN-based SI scheme is compared against DWT, DCT, and Adaptive Video Coding (VC) methods (Root Mean Square Error). Tables 2, 3, and 4 show the power consumption and time values.

A. PSNR Comparison

The PSNR is computed using the formula to evaluate the actual and decoded frames.

\[
PSNR = 10 \log \frac{255^2}{MSE}
\]

Figure 8. Restored image from FDCT

Fig.8.Restored image from FDCT

Fig.9.PSNR Performance with Respect No of Frames (Dataset 1)

Fig.10.PSNR Performance with Respect No of Frames (Dataset 2)

The PSNR of the presented FDCT with STN approach is comparable to prior methods in Figure 11. The number of frames is measured on the x-axis, while PSNR is measured on the y-axis. PSNR (in dB) with the suggested strategy is much greater than that of DWT, DCT, Adaptive VC schemes in the majority of frames.
The suggested FDCT with STN approach obtains 42.16 dB for the tennis spot dataset, whereas previous techniques including such as DWT, DCT, and Adaptive VC achieve 32.15dB, 35.12dB, and 42.15dB for 100 frames, respectively. The results show that the suggested approach is superior because it produces superior quality SIs for the video surveillance videos.

**Table 1. PSNR Comparison**

<table>
<thead>
<tr>
<th>No of frames</th>
<th>Dataset 1 (foreman)</th>
<th>Dataset 2 (tennis spot)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DWT</td>
<td>DCT</td>
</tr>
<tr>
<td>20</td>
<td>30.25</td>
<td>31.65</td>
</tr>
<tr>
<td>40</td>
<td>32.12</td>
<td>33.98</td>
</tr>
<tr>
<td>60</td>
<td>34.58</td>
<td>36.22</td>
</tr>
<tr>
<td>80</td>
<td>36.21</td>
<td>37.45</td>
</tr>
<tr>
<td>100</td>
<td>38.79</td>
<td>39.99</td>
</tr>
</tbody>
</table>

**B. MSE comparison**

The MSE is the difference between both the compression and original image's accumulated squares of the errors.

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2
\]  

(17)

![Fig.11. MSE for dataset 1](image-url)
The MSE evaluation of the proposed FDCT-STN with conventional DWT, DCT, and Adaptive VC approaches is shown in Figures 11 and 12. The x-axis reflects the amount of frames, while the y-axis represents the MSE. The calculated approximation coefficients for the present frames are generated using FDCT-STN in this study.

Table 2. MSE comparison

<table>
<thead>
<tr>
<th>No of frames</th>
<th>Dataset 1 (foreman)</th>
<th>Dataset 2 (tennis spot)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DWT</td>
<td>DCT</td>
</tr>
<tr>
<td>20</td>
<td>8.1</td>
<td>10.12</td>
</tr>
<tr>
<td>40</td>
<td>8.25</td>
<td>10.25</td>
</tr>
<tr>
<td>60</td>
<td>8.36</td>
<td>10.26</td>
</tr>
<tr>
<td>80</td>
<td>8.39</td>
<td>10.56</td>
</tr>
<tr>
<td>100</td>
<td>8.45</td>
<td>10.75</td>
</tr>
</tbody>
</table>

4.3 RMSE comparison

The proposed FDCT-STN method's RMSE is compared to the existing DWT, DCT, and Adaptive VC methods in Figure 13. The efficiency is enhanced by using an optimization strategy to choose the constant. The x-axis represents the number of frames, while the y-axis represents the RMSE. The suggested system obtains 1.5 RMSE for the tennis sport dataset, while other approaches including such DWT, DCT, and Adaptive VC achieve 2.5, 3.5, and 3.9 for 100 frames, correspondingly, according to experimental tests. The RMSE for dataset 1 and dataset 2 are shown in Figures 13 and figure 14. Table 4 compares the proposed HMO technique to previous techniques such as CSO and GA algorithms in terms of optimisation effectiveness. The optimisation techniques' reliability is increased for the amount of iteration, which implies that every type of input can be handled at any modality.
Fig. 13. RMSE for dataset 1

Table 3. RMSE comparison

<table>
<thead>
<tr>
<th>No of frames</th>
<th>DWT</th>
<th>DCT</th>
<th>Adaptive VC</th>
<th>FDCT-STN</th>
<th>DWT</th>
<th>DCT</th>
<th>Adaptive VC</th>
<th>FDCT-STN</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.83</td>
<td>2.72</td>
<td>2.64</td>
<td>1.5</td>
<td>3.01</td>
<td>2.85</td>
<td>2.79</td>
<td>1.45</td>
</tr>
<tr>
<td>40</td>
<td>2.96</td>
<td>2.82</td>
<td>2.75</td>
<td>1.52</td>
<td>3.12</td>
<td>2.99</td>
<td>2.85</td>
<td>1.52</td>
</tr>
<tr>
<td>60</td>
<td>3.00</td>
<td>2.84</td>
<td>2.81</td>
<td>1.56</td>
<td>3.17</td>
<td>3.01</td>
<td>2.93</td>
<td>1.53</td>
</tr>
<tr>
<td>80</td>
<td>3.11</td>
<td>2.98</td>
<td>2.87</td>
<td>1.58</td>
<td>3.25</td>
<td>3.14</td>
<td>2.99</td>
<td>1.54</td>
</tr>
<tr>
<td>100</td>
<td>3.21</td>
<td>3.04</td>
<td>2.94</td>
<td>1.59</td>
<td>3.22</td>
<td>3.28</td>
<td>3.14</td>
<td>1.56</td>
</tr>
</tbody>
</table>
Table 4. Optimization Performance

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Optimization Time (sec)</th>
<th>Total Generations (iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>CSO 112</td>
<td>CSO 75</td>
</tr>
<tr>
<td></td>
<td>GA 95</td>
<td>GA 65</td>
</tr>
<tr>
<td></td>
<td>HMO 85</td>
<td>HMO 70</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>CSO 154</td>
<td>CSO 75</td>
</tr>
<tr>
<td></td>
<td>GA 145</td>
<td>GA 70</td>
</tr>
<tr>
<td></td>
<td>HMO 140</td>
<td>HMO 68</td>
</tr>
</tbody>
</table>

D. Power Consumption Comparison

Power Consumption is the amount of energy consumed in a particular time period, which can be obtained as a numerical value:

\[ P = \frac{E}{t} \]

Where \( P \) denotes the power consumption, \( E \) denotes energy consumed by the device, and \( t \) time taken for an action.

Fig. 15 describes the power consumption rate for the proposed as well as existing approaches such as DWT, DCT, Adaptive VC, and FDCT-STN. As compared to the proposed approach with the existing approaches, the performance of the existing approaches are poor in terms of obtained results for dataset 1 and dataset 2.

E. Time Comparison

It represents how long the device takes to perform a particular work. The simulation time is the time taken by the network to complete the process of encoding and decoding of videos. Fig 15 illustrates the performance of the proposed approach with the existing approaches.
Fig.16. Time

Fig describes the time (s) for the proposed as well as the existing approaches such as DWT, DCT, Adaptive VC, approaches. As compared to the proposed approach with the existing approaches, the performance of the existing approaches are poor in terms of obtained results for dataset 1 and dataset 2.

5. CONCLUSION

To calculate SIs in DVCs, the suggested method used a Fast Discrete Curvelet Transform technique. The level-3 FDCT approach is used to mine approximation coefficients in this paper. STN is used to obtain approximation coefficients that are estimated. In this paper, comprehensive coefficients optimization is done using the HMO technique. IFDCT is utilised in the last stage to get approximated SIs in the wavelet coefficients. The SI-quality of this proposed work has indeed been developed to increase codec efficiency. Furthermore, as comparing to certain other systems, the experimental data of the suggested scheme’s simulations achieved higher performance measure results in MSE, PSNR, Power Consumption, Time, and RMSE. As comparing against current solutions, the proposed approach has outperformed them. In future, this work extends into the lightweight artificial intelligence approach for the resource constrained based video transmission devices and services. Furthermore, security is added to improve the sensitivity of the video while transmitting from the source node.

DECLARATION:

Ethics Approval and Consent to Participate:
No participation of humans takes place in this implementation process

Human and Animal Rights:
No violation of Human and Animal Rights is involved.

Funding:
No funding is involved in this work.

Conflict of Interest:
Conflict of Interest is not applicable in this work.

Authorship contributions:
There is no authorship contribution

Acknowledgement:
There is no acknowledgement involved in this work.

REFERENCES


