Ratio Mathematica Volume 48, 2023

Deep Features and Clustering Based Keyframes Selection with Security

Prachi Chauhan*
Hardwari Lal Mandoria†
Alok Negi‡

Abstract

Multimedia processing and distribution have become vulnerable due to vital information's enormous quantity and significance. Therefore, comprehensive technologies and algorithms are required to transmit messages, images, and video files safely. This paper proposes a secure framework by acute integration of video summarization and image encryption. Three parts comprise the proposed cryptosystem framework. Firstly, the useful, informative frames are extracted using an efficient and lightweight technique that uses the color histogram-clustering (RGB-HSV) approach's processing capabilities. Each video frame is represented by deep features based on an enhanced pre-trained Inception-v3 network. After that summary is obtained using the K-means optimal clustering algorithm, the representative key frames are extracted using the cluster's highest possible entropy nodes. Experimental validation on two well-known standard datasets demonstrates the proposed methods' superiority to numerous state-of-the-art approaches. Finally, the proposed framework performs an efficient image encryption and decryption algorithm by employing a general linear group function GL_n(F). The analysis and testing outcomes prove the superiority of the proposed adaptive RSA.

Keywords: Clustering; Color Histogram; General Linear Group; Image Encryption and Decryption; Inception-V3; Video Summarization

*Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India; 49422 prachichauhan@gbpuat-tech.ac.in

†Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India; dr-mandoria1@gmail.com

‡UPES, Dehradun, Uttarakhand, India; alok.negi@ddn.upes.ac.in

[1]Received on February 2, 2023. Accepted on July 8, 2023. Published on October 15, 2023. DOI: 10.23755/rm.v41i0.1139. ISSN: 1592-7415. eISSN: 2282-8214.

1 Introduction

Multimedia content is expanding exponentially in the digital era and has taken on significant importance in formulating modern life's methods. Processing mechanisms in this situation have high operating costs and limited resources. Every firm needs videos encompassing social networks, web conferencing, adaptive traffic control, environmental monitoring, and security monitoring. Video data must be processed significantly to recognize the instructive objects and scenes. On the other hand, the storage and trans- mission of data alongside security has grown to be a significant concern. The system creates enormous amounts of data, which must be processed securely and in real time. Thus, a few academics presented various strategies, including video summaries and cryptography methods, to overcome computational complexity and security Muham- mad et al. [2020], Hamza et al. [2019].

Video summarization (VS) is extracting significant and brief video clips that accurately convey the entire overarching storyline of the original video Sahu and Chowdhury [2020]. The generated summary must have two essential characteristics: it must always include high-priority information and be devoid of duplication scenes. As a result, the demand for automated and effective image retrieval solutions is rising. These technologies are designed to identify redundant images Rani and Kumar [2020] automatically. A traditional image retrieval method compares the feature vectors of different images based on various distance metrics to assess how similar the images are. The most com- parable images will be retrieved once a query image feature vector is compared to those in the database. The first and most prominent visual feature in image retrieval and indexing is the color feature da Silva Torres and Falcao [2006]. The ability to demonstrate visual content in images, the simplicity and high reliability of extracting color information from the images, the relative strength in separating images from one another, the relative robustness to background complexity, and independence from image size and orientation are the most significant benefits of color feature Alamdar and Keyvanpour [2011]. One of the most extensively used color feature extractors is the color histogram because it is easily implemented, computationally efficient, and invariant to rotation and slight changes in viewing position Li and Jiang [2016].

The keyframes extracted represent the vital scene in the original video. However, subject to confidentiality, integrity, authenticity, and other factors, such information is not safe to share online. So, it is strongly advised to secure frames Muhammad et al. [2020], Hamza et al. [2019], Sixing et al. [2012], Mukhedkar et al. [2015], Ramanujam and Karuppiah [2011], Zia et al. [2022]. As a result, the focus is applying the security algorithm to solve these kinds of issues in our work. The innovative concept uses a linear group to make the key space more difficult. It consists of exponential modulus as a function of a group for the encryption network and hides all original image features. The decryption procedure, which is the inverse operation of the encryption procedure, is comparable to conventional encryption-decryption techniques.

The rest of the paper is organized as follows: Theoretical information related to this research is discussed in section 2. Section 3 proposes a solution to the current challenge in secure video summarization, while Section 4 outlines the experiments with a comparison of proposed and existing methods. Section V concludes with conclusions and future scope.

2 Related Work

The multimedia device intends to automatically perform an activity or provide real-time information from the data centers. In other words, the critical goal is to collect accurate, intelligent data that performs via device mastery approaches under extraordinary circumstances. Numerous different kinds of research have been done and are still being pursued for this goal. A technique to directly measure pixel differences between two frames was introduced by Wu and Xu [2013]. Based on creating differentiable histograms and histogram-based loss functions, Avi-Aharon et al. [2020] introduced the DeepHist, a unique deep learning approach for image-to-image translation.

To address the shortcomings of existing methodologies that neglect to consider variations in shot complexity, Liang et al. [2021] proposed a news report summarization scheme based on SURF features and an enhanced clustering algorithm. The shot's abrupt and gradual boundaries were detected using SURF features. The color histogram of the video frames within the shot was then clustered using an enhanced clustering algorithm. When tested on news video datasets, the proposed approach recorded an average accuracy of 93.33 percent and a recall rate of 97.22 percent in shot boundary detection.

In addition, Gygli [2018] proposed an FCNN to discover shot detection end-to-end, from pixels to final shot boundaries, and enabled to use of a large temporal context without any need to repetitively sequence frames. Sou**c**ek and Loko**c** [2020] presented TransNet-V2, a deep network for identifying prevalent shot transitions, that further represented a significant preliminary step of video analysis procedures. The computer vision and multimedia industries are now interested in researching egocentric video summarization. Gygli et al. [2014] segmented the clip into segments based on motion cues and gave each section a score. Finally, the best segments from the summary were chosen within time limits.

To solve the abovementioned issues, Kumar and Shrimankar [2017] developed a local-alignment-based FASTA strategy to summarize the events in multi-view videos. The FASTA algorithm was then used to capture interview dependencies between different video views via local alignment before object tracking that extracted the frames with low activity. The presented work highlights some well-known research on clustering- based video summarization (Sahu and Chowdhury [2020], Rani and Kumar [2020]) because that is the basis of our solution.

The most widely used clustering-based approaches are as follows: Sahu and Chowdhury [2020] introduced CSMIK K-means, which was K-means based on a center-surroundmodel (CSM) and an Integer Knapsack type formulation (IK). CSMIK K-Means vali- dated various cluster groupings and calculated the optimal number of clusters and the associated summary. In the continuation, authors Rani and Kumar [2020] presented a keyframe extraction technique focused on four visual features and a clustering analysis based on the Kohonen Self-Organizing Map to obtain the most prominent frames from the list of frames generated after fusion. Papadopoulos et al. [2013] proposed another fully automated VS method in which a dynamic calculation process calculates the final number of clusters for summarization.

Due to the rapid development of digital video technology, much data is produced through video calls or internet video conferencing, for which participants set up an active session for contact with one another. Web video conferencing, used in online seminars or webinars, means the sender sends their information online. As a result, it is necessary to encrypt confidential information using an effective cryptographic technique before outsourcing. In a protected digital world, the ciphertext pair and the integrity of the ciphertext pair data significantly impact network efficiency. To obtain a strong security statement, it must withstand all known cryptanalysis.

Against Simon and Simeck, authors Lu et al. [2022] developed neural distinguishers (NDs) and related key neural distinguishers (RKNDs). Simon32/64's ND and RKND achieved 11- and 11-round accuracy of 59.55 percent and 97.90 percent, respectively. In 13 rounds of Simon64/128, the ND achieved an accuracy of 60.32 percent, while the RKND achieved an accuracy of 95.49 percent. In order to continue ensuring the secure and efficient transmission of video data.

According to Cheng et al. [2020], the essential semantic elements (IPM, MVD, residual coefficient, and delta QP) were encrypted in each slice. This method used the segmentation capabilities of the H.264 encoding, and the procedure of selective video encryption was based on video coding technology and a four-dimensional hyperchaotic scheme. The authors also examined the perceived quality of encrypted video using several reference video sequences with motion, texture, and objects. Further, to ensure image security, Gaherwar et al. [2022] presented a method for safeguarding images by selective alteration (SISA). The technique used selective encryption or blurring to encrypt an image, which reduced processing time without compromising security.

By applying deep learning techniques, authors Ding et al. [2020] suggested a method for encrypting and decrypting medical images (specifically DLEDNet). To encrypt and decrypt the medical image, the Cycle-GAN network was chosen as the learning network. A target domain directed the technique for modeling to implement the encryption procedure. Additionally, an ROI-mining network was developed to extract the ROI from medical images that had been encrypted, allowing DLEDNet to segment the relevant organs or tissues in ciphertext environments without having first to decrypt the images.

Instead of the normal video acquired from a third-person view, the egocentric video, often captured by first-person equipment, presents several issues. Since the camera is constantly moving, the identical items could disappear and reappear in succeeding frames in an egocentric clip. Thus, creating automated, instructive summaries from egocentric videos has become a tough challenge. This study introduces the best clustering method for egocentric video summaries. The goal of the solution workflow is to obtain better accuracy while using minimal processing resources. However, specific solutions need more robustness when it comes to securing keyframes. Another issue is a mathematical transformation for noise and cropping-resistant image encryption. Consequently, additional in-depth evaluations to choose the best cryptosystem are always necessary.

3 Proposed Methodology

The four key stages of our proposed framework are as follows: a) Preprocessing; b) Feature extraction using the modified Inception model and obtaining a set of several clusters (k values) for key-frame selection; c) Key-frame security. An overview of our proposed solution is described as a block diagram in figure 1.

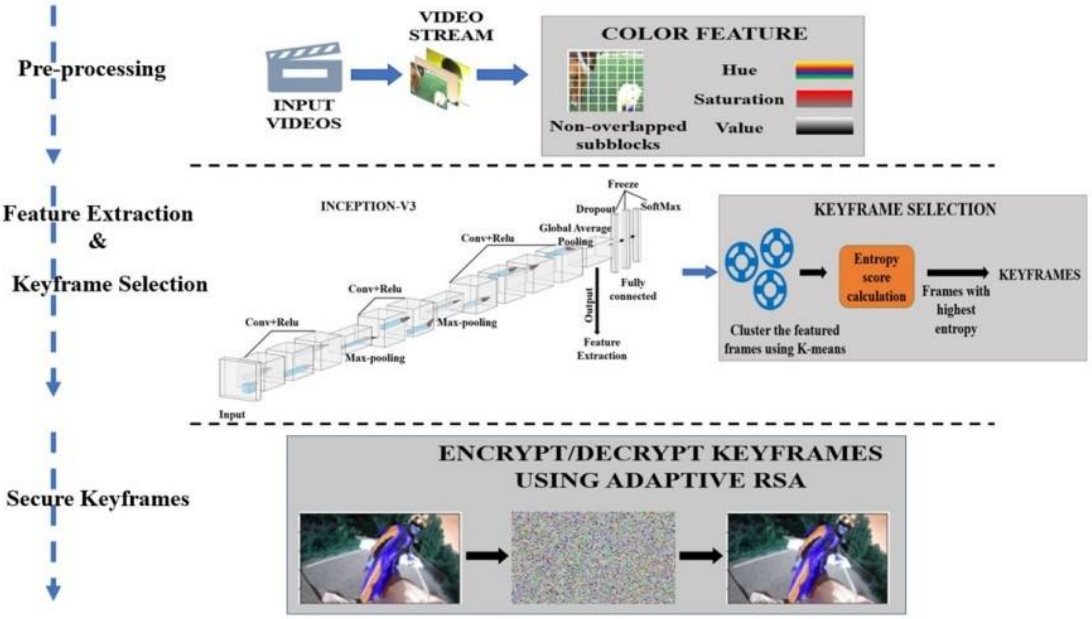


Figure 1: Block Diagram of Proposed Methodology

3.1 Pre-processing

The video is comprised of multiple frames. Therefore, choosing a precise number of frames from a sequence of videos at fixed intervals is necessary for preprocessing. Hence, the video sequence V_{id} is the set of several frames F_N at a time 't', represented by equation (1), in which F(t+i) refers to a frame at a time (t+i).

$$V_{id} = \sum_{i=1}^{N} F(t+i) \tag{1}$$

The size of the V_{id} is extensive as it contains useful and useless frames, which affects the computational time. Eliminating useless frames from V_{id} is mandatory. Hence, a global color histogram is used to analyze statistical color frequency in an image by solving the problems of change in translation, rotation, and angle of view Srivastava et al. [2015]. For histogram computation, the frame is divided into three color channels (R, G, B), and each color channel is further split into non-overlapping sub-blocks. Let F(x),F(x+1),F(x+3),. ,F_N are the frames of dimension P*Q. The frames are further divided into several sub-blocks S_B with dimension M*N, where $M < P$ and $N < Q$. The histogram difference of every channel is measured by histogram intersection as given in equation (2-4).

$$H_R(F(x), F(x+i)) = \frac{1}{S_B} \sum_{k=1}^{S_B} \left[1 - \sum_{i=1}^{24} min \left(H_{x,k}(l)_R, H_{x+1,k}(l)_R \right) \right] \tag{2}$$

$$H_G(F(x), F(x+i)) = \frac{1}{S_B} \sum_{k=1}^{S_B} \left[1 - \sum_{i=1}^{24} min \left(H_{x,k}(l)_G, H_{x+1,k}(l)_G \right) \right] \tag{3}$$

$$H_B(F(x), F(x+i)) = \frac{1}{S_B} \sum_{k=1}^{S_B} \left[1 - \sum_{i=1}^{24} min \left(H_{x,k}(l)_B, H_{x+1,k}(l)_B \right) \right] \tag{4}$$

where, $(H_{x,k}(l))$, $(H_{x+1,k}(l))$ are the color histogram of k_{th} section of three RGB channels of frame F(x), F(x+1) respectively.

Finally, the average of all channel differences, D_H is computed as shown in equations (5), and it is compared to the input threshold T_H in function (6) for the eliminationof useless frames.

$$D_{H-RGB} = \frac{\sum (H_R) + (H_G) + (H_B)}{3} \tag{5}$$

$$\{D_{H-RGB} < T_H\} \tag{6}$$

If the above condition satisfies, then frames are considered beneficial and informative.

3.2 Feature and Keyframe Extraction

To obtain visual descriptors, the presented method uses Inception-V3 (Szegedy et al. [2016]). In comparison to dppLSTM (Zhang et al. [2016]), M-AVS (Ji et al. [2019]), GoogleNet (Sahu and Chowdhury [2020]), and SUM/GAN (Mahasseni et al. [2017]), the model has the best feature to decrease the number of resources and enhance speed. The fine-tuning of all the convolutional layers till the global average pooling at the end of the network occurs in this work. The final three layers are removed, and the features generated by the bottleneck layer are used. After that, the well-known k-means algorithm (Ahmed et al. [2020]) is applied to cluster featured frames. K-means separates the area of focus from the background and utilizes a color palette to represent the visual frame as the human eye would perceive it.

The number of clusters to be set is 15 percent and 20 percent for SUMME and TVSUM datasets, respectively. After that, the frame with the highest channel entropy in the cluster is declared as a keyframe. The entropy being discussed here is Shannon Entropy, which quantifies the informational content of an image and specifies how much ambiguity or randomness there is in an image. The information that an image contains can be calculated using equation (7) to determine its quality, in which 'k' is the color space of an image and 'p' is the ratio of the number of occurrences of intensity level to total pixels.

$$H(I) = \sum_{i \in I} p_i(i) * \log \frac{1}{p_i(i)} \tag{7}$$

where p_i = p(k=i) ; i=1,2,3,...,'m'

Since then, the entropy computed from channels and HSV for hue, saturation, and intensity value has been used for the color histogram because it effectively divides RGB into luminosity and chromaticity Nazir et al. [2018] and simulates how humans perceive color using formulas (8-10).

$$H = \cos^{-1} \frac{\frac{1}{2}(R-G) + (R-B)}{\sqrt{(R-G)^2 - (G-B)(R-B)}} \tag{8}$$

$$S = 1 - \frac{3\,[min(R, G, B)]}{R + G + B} \tag{9}$$

$$V = \frac{R + G + B}{3} \tag{10}$$

Where 'H' is the hue dominant wavelength of color in image collection, 'S' is the saturation magnitude of white light within the image, and 'V' is the intensity value or lightness that describes how dark or light a color is. Each RGB pixel in a colorful image is 3 bytes, with each color having an intensity range of 0 to 255 and a total of

256*256*256 colors that can be displayed. As a result, information entropy for frame F of size M*N with RGB channel account of A_{RGB} is given in equation (11).

$$F_{RGB} = -\log_2 256^{M*N*A_{RGB}} \tag{11}$$

3.3 Keyframes Security

The unique characteristics of video, such as its enormous size and volume, make it impossible to ensure video security, and most currently used security methods need help to handle complex data in real-time. Therefore, an adaptive RSA encryption and decryption algorithm is proposed in this research as given in algorithm 1. The ARSA uses three prime numbers because multiplying three significant prime numbers quickly yields the desired result, but doing the opposite requires a lot of computing power. The exponential modulus (g) is taken as the number of elements in the general linear group GL_n(F). Under matrix multiplication, the GL_n(F) is a set of integrated n*n matrices with elements in finite field F. The number of elements in $GL_n(F)$ is $\prod_{k=0}^{n-1}(p^n - p^k)$. Suppose there are n*n matrices with linearly independent rows. The first row can be anything other than the zero rows, so there are p^n-1 possibilities. The second row must be linearly independent of the first, so there are p^n-p possibilities. The i^{th} row also be linearly independent from the first i 1 rows, so total possibilities are $p^n\ p^{i-1}$. Therefore, $(p^n - 1).(p^n - p)\ p^n - p^{i-1}$ equals to $\prod_{k=0}^{n-1}\ p^n - p^k$.

Algorithm : Adaptive RSA

---

1. Choose three prime numbers p,q, and r randomly and independently of each other.
2. Calculate : n = p*q*r and g = $g_p + g_q + g_r$
   (i) $g_p = (p^2 - 1) . (p^2 - p)$
   (ii) $g_q = (q^2 - 1) . (q^2 - q)$
   (iii) $g_r = (r^2 - 1) . (r^2 - r)$
3. Choose an integer e, $1 < e <$ g , such that gcd(e,g) = 1
4. Compute the private exponent d, $1 < d <$ g , such that ed ≡ 1 mod g
5. Encryption: 'M' is the original RGB image vector
   C = M^e(mod n)
6. Decryption: 'C' is the encrypted RGB image vector
   M = C^d(mod n)

4 Experimental Results

4.1 Datasets Description

The SUMME and TVSUM datasets were tested to validate the method's effectiveness. The TVSUM provides 50 videos from various genres, including vlogs, news, and egocentrics, although the SUMME has 25 videos in the sports, holidays, and events categories. Table 1 summarizes the characteristics of both datasets.

Dataset	Number of clips	Resolution	Shots	Size
SUMME	25	720*1280	390	1hour 10 minutes
TVSUM50	50	360*540	1720	3hours 50 minutes

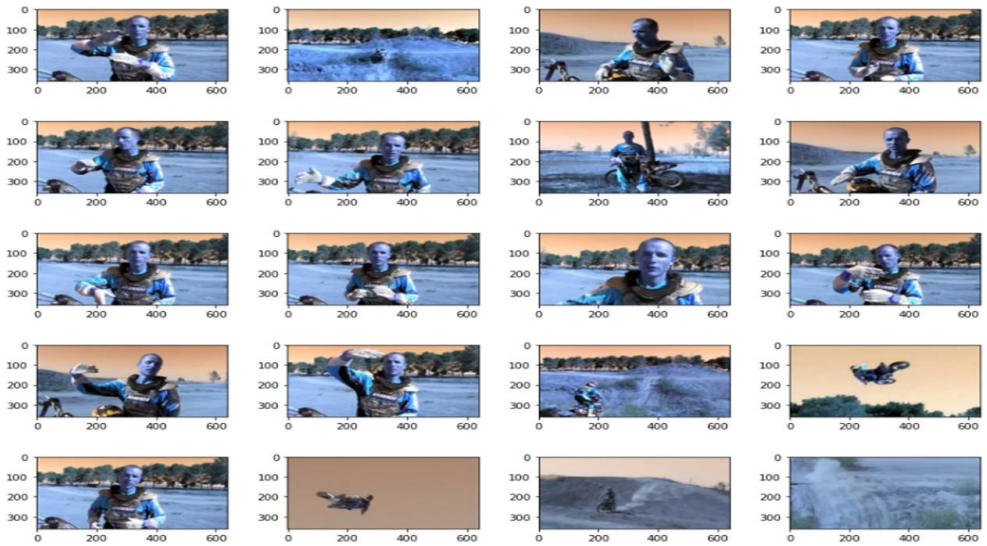
Table 1: Characteristics Details of Datasets

In the SUMME and TVSUM databases, the maximum number of user summaries (Sahu and Chowdhury [2020]) per video is 15 and 20, respectively. Figure 2 shows some extracted key-frames of the clips of scuba from SUMME and bike-tricks, grooming animals from TVSUM datasets. There are no redundant frames in the extracted key-frames, which accurately reflect most of the information in the video.

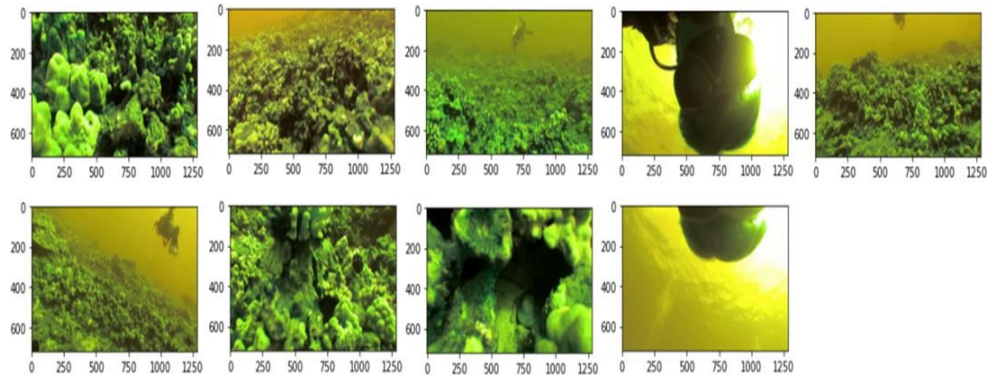
4.2 Qualitative Analysis

User summary for summarization from Sahu and Chowdhury [2020] and Zhang et al. [2020] are considered the ground truth for both experiment datasets. This work used the scripts and other helpful information Sahu and Chowdhury [2020] to get the findings and summaries. Another potentially valuable work was completed by Zhang et al. Zhang et al. [2020], who used the SUMME database to generate their outcomes and summary. The scripts and other beneficial information are obtained, such as reviewing the generated frames for video highlights since the strategy is reliable for generating frames.

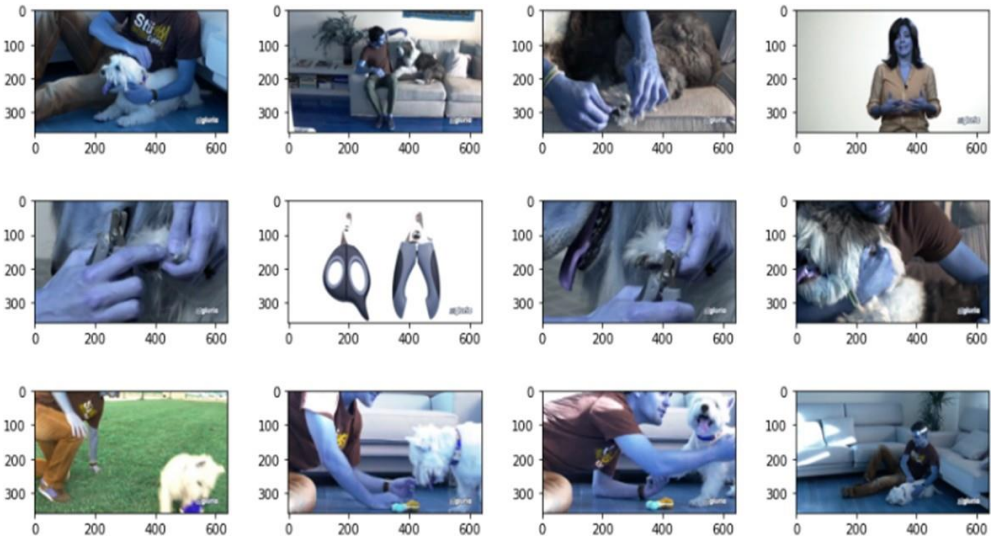
The qualitative evaluation results are displayed in figure 3, and these clearly demonstrate that the proposed technique can summarize key-frames that contain representative objects, such as the instant a person is jumping or flying. On the other hand, the CSMIK k-means and online motion techniques are more likely to choose frames with limited in- formation. The same is accurate for the bike-polo key-frames of the proposed method given in figure 4 as compared to Gygli [2018], which includes every action such as setting up the goal, giving the ball to a player, receiving a shot, tracking a player hoppinginto position, and quickly crossing a player.



(a) Bike-Tricks



(b) Scuba



(c) Grooming animals

Figure 2: Few frames of summary generated by proposed approach

Figure 3: Representative frames of Base-jumping category of SUMME dataset: 1st row (Proposed framework), 2nd row Sahu and Chowdhury [2020] and 3rd row Zhang et al. [2020]

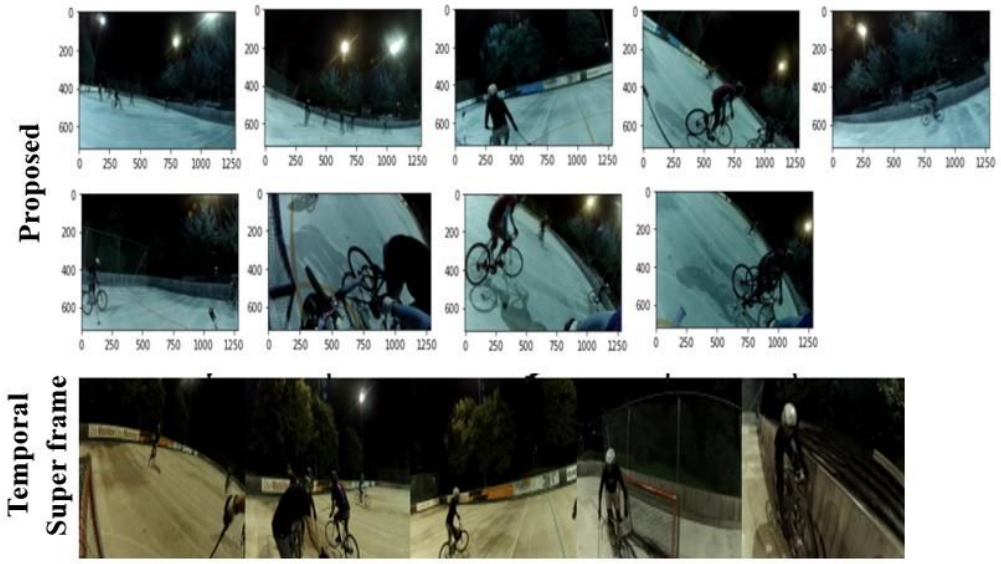


Figure 4: Representative frames of summarized events Bike-polo: 1st row (Proposed framework), 2nd row Gygli [2018]

4.3 Quantitative Analysis

Recall, precision, and f1-score are metrics for performance evaluation. The term ”precision” relates to a method’s accuracy, which can be determined by counting the improperly extracted key-frames. The recall value refers to the possibility of each key-frame existing in the ground truth. In equation (12), $KF_{matched}$ indicates how many key-frames were matched to the ground truth, and $KF_{extracted}$ indicates how many key-frames were extracted overall using our method. The number of key-frames in the ground truth summary is denoted by $KF_{groundtruth}$ in equation (13), and the f1-score is computed via equation (14).

$$Precision = \frac{KF_{matched}}{KF_{extracted}} \tag{12}$$

$$Recall = \frac{KF_{matched}}{KF_{groundtruth}} \tag{13}$$

$$F1\text{-}\ score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{14}$$

The analysis of recall and precision under various types of videos in a comparison Table 2 with the algorithms by Liang et al. [2021] and TransNetV2 Sou**c**ek and Lokoc [2020], and Table 3 below displays the f-score comparison results with Sahu and Chowdhury [2020], Huang and Wang [2019], Gygli [2018], Zhao et al. [2018], Khan et al. [2019].

The proposed approach recorded 87.50 and 93.33 percent precision and recall scores for the SUMME dataset as shown in Table 2 while it recorded 90.47 and 95.04 percent for the TVSUM dataset. F1 score which provides the balance between precision and recall is recorded at 89 and 92 percent for the SUMME and TVSUM datasets respectively as shown in Table 3. The comparison results confirm the effectiveness of the algorithm.

Methods	**SUMME**		**TVSUM**	
	Precision	**Recall**	**Precision**	**Recall**
SURF Liang et al. [2021]	-	-	87.91	95.93
TransNetV2 Soucek and Loko**c** [2020]	-	-	92.02	92.56
Proposed	**87.50**	**93.33**	**90.47**	**95.04**

Table 2: Quantitative Evaluation of Precision and Recall

Methods	**SUMME**	**TVSUM**
CNN+CSMIK K-means Sahu and Chowdhury [2020]	0.452	0.629
CapsNet Huang and Wang [2019]	0.89	0.87
FCNN Gygli [2018]	0.87	0.85
HSA-RNN Zhao et al. [2018]	0.89	-
CNN+LSTM Khan et al. [2019]	-	0.84
Proposed	**0.89**	**0.92**

Table 3: Quantitative Evaluation of F1-score

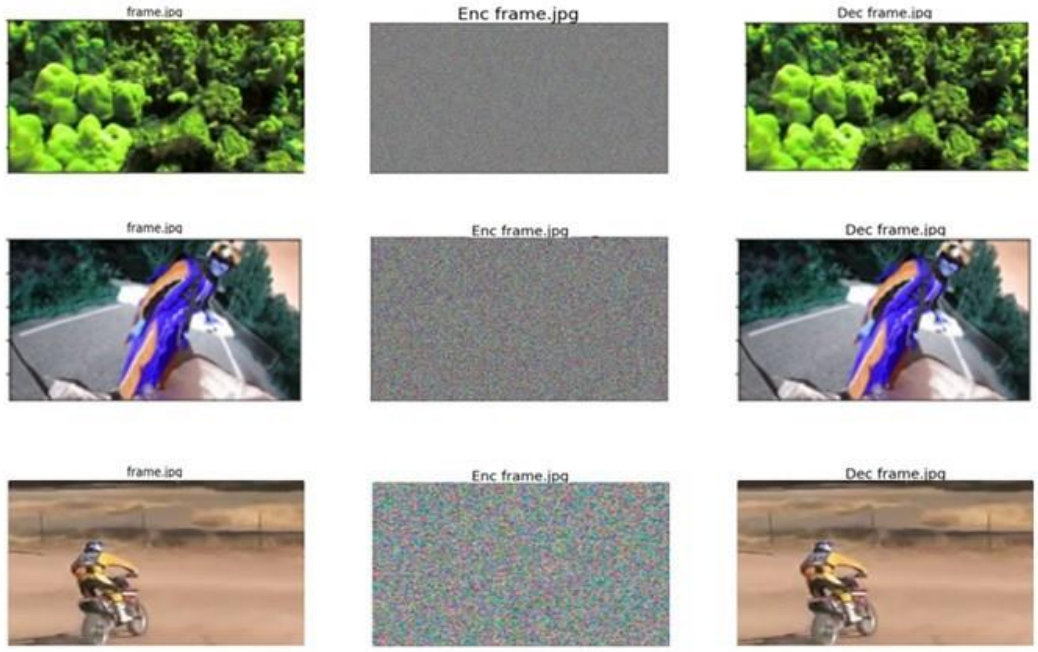


Figure 5: Encryption and Decryption of few representative key-frames using adaptive RSA

4.4 Security Analysis

Key-frames encryption and decryption are performed using general linear group GL_n(F) algorithms to withstand extensive attacks. Here, GL_n(F) of degree' n' over prime numbers is the set of n*n convertible matrices with entries from prime numbers. The RGB channel of the image determines the group degree, with n set to 3. Two keys were created using adaptive asymmetric encryption's RSA exponential modulus function. These keys can invert each other's encrypted data but cannot decrypt their data.

The technique offers a high degree of protection for the key-frames by hiding all features of the original key-frames, as shown in figure 5. The technique is highly vulnerable to the secret key as a value obtained by GL_n(F) function. As a result, the pro- posed image cryptosystem does not provide hackers with any helpful information, thusefficiently validating security.

5 Conclusion and Future work

The digital world is rapidly evolving, making multimedia processing and distribution vulnerable. A significant portion of redundant video data is produced due to recent improvements in digital networks. Its management, analysis, and transmission are com- plex. Hence, there is a need for image prioritization. This study first extracts the informative frames using a compelling video summarizing technique. A secure framework is proposed by integrating video summarization and image encryption. The framework extracts informative frames using RGB-HSV, Inception-v3 network deep features, and K-means optimal clustering. The framework also performs efficient image encryption and decryption using GL_n(F) function.

The experimental results support our algorithm's superior precision, recall, and f1-score metrics compared to existing state-of-the-art approaches. The security and privacy of the retrieved key frames during communication are most important because these are necessary for subsequent analysis. As a result, proposed adaptive RSA to encrypt key-frames before transmission. Here, the exponential modulus that assists in creating two keys is calculated using the general linear group. The algorithm provides high safety as it hides all the featured information. The key drawback of the proposed method-ology is that it keeps the visual representation of the decrypted data while encrypting numerous frames at once rather than one-to-one. Therefore, to increase the overall system's security, future development will focus on dynamic keys rather than traditional encryption keys. Experimental validation on standard datasets shows its superiority to state-of-the-art approaches.

References

M. Ahmed, R. Seraj, and S. M. S. Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

F. Alamdar and M. Keyvanpour. A new color feature extraction method based on quad-histogram. *Procedia Environmental Sciences*, 10:777–783, 2011.

M. Avi-Aharon, A. Arbelle, and T. R. Raviv. Deephist: Differentiable joint and color

histogram layers for image-to-image translation. *arXiv preprint arXiv:2005.03995*, 2020.

S. Cheng, L. Wang, N. Ao, and Q. Han. A selective video encryption scheme based on coding characteristics. *Symmetry*, 12(3):332, 2020.

R. da Silva Torres and A. X. Falcao. Content-based image retrieval: theory and applications. *RITA*, 13(2):161–185, 2006.

Y. Ding, G. Wu, D. Chen, N. Zhang, L. Gong, M. Cao, and Z. Qin. Deepedn: a deep-learning-based image encryption and decryption network for internet of med- ical things. *IEEE Internet of Things Journal*, 8(3):1504–1518, 2020.

P. Gaherwar, S. Joshi, R. Joshi, and R. Khengare. Sisa: Securing images by selective alteration. In *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, pages 729–740. Springer, 2022.

M. Gygli. Ridiculously fast shot boundary detection with fully convolutional neural networks. In *2018 International conference on content-based multimedia indexing (CBMI)*, pages 1–4. IEEE, 2018.

M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014.

R. Hamza, A. Hassan, T. Huang, L. Ke, and H. Yan. An efficient cryptosystem for video surveillance in the internet of things environment. *Complexity*, 2019, 2019.

C. Huang and H. Wang. A novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):577–589, 2019.

Z. Ji, K. Xiong, Y. Pang, and X. Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717, 2019.

M. Z. Khan, S. Jabeen, S. ul Hassan, M. Hassan, and M. U. G. Khan. Video summarization using cnn and bidirectional lstm by utilizing scene boundary detection. In *2019 International conference on applied and engineering mathematics (ICAEM)*, pages 197–202. IEEE, 2019.

K. Kumar and D. D. Shrimankar. F-des: Fast and deep event summarization. *IEEE Transactions on Multimedia*, 20(2):323–334, 2017.

M. Li and X. Jiang. An improved algorithm based on color feature extraction for image retrieval. In *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 2, pages 281–285. IEEE, 2016.

B. Liang, N. Li, Z. He, Z. Wang, Y. Fu, and T. Lu. News video summarization combining surf and color histogram features. *Entropy*, 23(8):982, 2021.

J. Lu, G. Liu, Y. Liu, B. Sun, C. Li, and L. Liu. Improved neural distinguishers with (related-key) differentials: Applications in simon and simeck. *arXiv preprint arXiv:2201.03767*, 2022.

B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017.

K. Muhammad, T. Hussain, and S. W. Baik. Efficient cnn based summarization of surveillance videos for resource-constrained devices. *Pattern Recognition Letters*, 130:370–375, 2020.

M. Mukhedkar, P. Powar, and P. Gaikwad. Secure non real time image encryption algorithm development using cryptography & steganography. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE, 2015.

A. Nazir, R. Ashraf, T. Hamdani, and N. Ali. Content based image retrieval system by using hsv color histogram, discrete wavelet transform and edge histogram descriptor. In *2018 international conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–6. IEEE, 2018.

D. P. Papadopoulos, V. S. Kalogeiton, S. A. Chatzichristofis, and N. Papamarkos. Automatic summarization and annotation of videos with lack of metadata information. *Expert Systems with Applications*, 40(14):5765–5778, 2013.

S. Ramanujam and M. Karuppiah. Designing an algorithm with high avalanche effect. *IJCSNS International Journal of Computer Science and Network Security*, 11 (1):106–111, 2011.

S. Rani and M. Kumar. Social media video summarization using multi-visual features and kohnen's self organizing map. *Information Processing & Management*, 57(3): 102190, 2020.

A. Sahu and A. S. Chowdhury. Summarizing egocentric videos using deep features and optimal clustering. *Neurocomputing*, 398:209–221, 2020.

X. Sixing, S. Xin, L. Bing, et al. New image encryption technology of image based on computer generated hologram [j]. *Laser & Optoelectronics Progress*, 49(4):040902, 2012.

T. Souček and J. Lokoč. Transnet v2: an effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.

D. Srivastava, R. Wadhvani, and M. Gyanchandani. A review: color feature extraction methods for content based image retrieval. *International Journal of Computational Engineering & Management*, 18(3):9–13, 2015.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Z. Wu and P. Xu. Shot boundary detection in video retrieval. In *2013 IEEE 4th International Conference on Electronics Information and Emergency Communication*, pages 86–89. IEEE, 2013.

K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016.

Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing. Unsupervised object-level video summarization with online motion auto-encoder. *Pattern Recognition Letters*, 130: 376–385, 2020.

B. Zhao, X. Li, and X. Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7405–7414, 2018.

U. Zia, M. McCartney, B. Scotney, J. Martinez, M. AbuTair, J. Memon, and A. Sajjad. Survey on image encryption techniques using chaotic maps in spatial, transform and spatiotemporal domains. *International Journal of Information Security*, pages 1–19, 2022.