Object Pre-processing using Motion Stabilization and Key Frame Extraction with Machine Learning Techniques

Kande Archana[1]
Research Scholar
Department of CSE
Jawaharlal Nehru Technological University (JNTU)
Hyderabad, India, Telangana State

V Kamakshi Prasad[2]
Professor
Department of CSE
Jawaharlal Nehru Technological University (JNTU)
Hyderabad, India, Telangana State

***Abstract*—Video information processing is one of the most important application areas in research and to solve in various pre-processing issues. The pre-processing issues such as unstable video frame rates or capture angle, noisy data and large size of the video data prevent the researchers to apply information retrieval or categorization algorithms. The video data itself plays a vital role in various areas. This work aims to solve the motion stabilization, noise reduction and key frame extraction, without losing the information and in reduced time. The work results into 66% reduction in key frame extraction and nearly 6 ns time for complete video data processing.**

***Keywords*—Information loss preventive; mean angle measure; key frame extraction; moving average; dynamic thresholding**

I. Introduction

Key-frame extraction from first-person vision (FPV) is a fundamental tool for highlighting and memorizing meaningful moments from a person's life. Selecting pivotal frames is challenging when using head-mounted FPV cameras due to the scene's inherent uncertainty. Because of the inherent instability of head-mounted cameras, FPV footage is noisier than TPV footage. In order to extract key frames, most algorithms consider TPV scenes to be static. Extraction of key frames from noisy first-person-view videos is still in its infancy. While human motion is ubiquitous in daily life, most key-frame fractionation algorithms rely on information from first-person videos. FPV videos rely heavily on motion capture integrated with visual scenes due to the rapid nature of scene transitions. The authors of this study propose a new key-frame extraction technique for FPV videos to attenuate background noise and identify potentially important events. Both the sparse-based and graph-based multi-sensor integration models proposed for key-frame extraction perform well on the shared space. Experiments with multiple datasets support the claim that the proposed key-frame extraction techniques enhance "extraction precision and coverage of entire video sequences" [1].

When applied to dynamic regions of interest, video SAR can enhance automatic retrieval of information. Extracting key frames from a video efficiently processes large amounts of data. Background subtraction using computer vision is suggested for automatic key-frame selection in Video SAR scattering. Key frames for scattering in Video SAR are revealed by a universal parameterization model. To differentiate the scattering key-frame state of transient persistence and disappearance, we propose the use of the "sub aperture energy gradient (SEG) and a modified statistical and knowledge-based object tracker (MSAKBOT)". Multitemporal video sequences are no match for the proposed SEG-MSAKBOT method, which evaluates pelting key frames in a more comprehensive and adaptable manner. Experimental findings and performance evaluation on two actual airborne Video SAR datasets with coherent integration angles [2].

VSUMM is widely used for processing large amounts of video information. Frames that best capture the essence of a video's subject are chosen by VSUMM. A novel framework is proposed for elucidating content and motion to solve these problems. A cross - functional and cross motion curve is created by Capsules Net, which has been taught to extract spatiotemporal information. Next, we propose an approach to automatically extract individual shots from continuous video feeds by analyzing their sensitivity to a variety of transition effects. Key-frame patterns in under shots can be selected using a self-attention model, key static images can be chosen for video content categorization, and optical movements can be calculated for video motion summarization. Results from experiments show that the method achieves competitive results [3].

An essential part of video retrieval is the extraction of key frames. For videos with multiple scenes and actions, the current state of the art in key frame extraction falls short. This article proposes a model for image saliency extraction that is aided by deep prior information. The paper proposes a saliency extraction algorithm to find important details in a given image. Here we describe a novel approach to finding the most important moments in a video. In this article, we build an image-integrated model of visual attention to track a combination of multiple bottom-level features and the skin color confidence map of the target. Extraction of the moving targets. The algorithm has been shown to effectively grasp pedestrian information in moving videos and provide samples of motion targets for post-processing [4] based on experimental results.

The rest of the work is organized such as in the Section – II, III and IV the foundational methods for motion stabilization, noise reduction and key frame extractions are discussed, further based on the baseline methods, the recent research improvements are discussed in the Section – V, further the persistent research problems are discussed in the Section – VI and based on the analysis of the problem, the proposed solutions are discussed in Section – VII and Section VIII, the obtained results from the proposed solutions are discussed in the Section – IX and again compared with the other benchmarked algorithms in Section – X. The research presents the conclusion in the Section – XI.

II. Foundational Method for Motion Stabilization

Firstly, in this section of the work, the baseline method for motion stabilization is analyzed. The process for motion stabilization is highly crucial as the unstable video data influences the decisions during video stabilization process.

Assuming that, the complete video data is V [] and each frame is denoted as f_i. Then, for n number of total frames, the relation can be formulated as,

$$V[] =< f_1, f_2, f_3 .... f_n > \tag{1}$$

Further, assuming that the function ϕ is an arbitrary function to extract the dimension, d_X[] and angle of video capture, a_X from the frames. Thus, this can be furnished as,

$$\phi\{ f_X \} \Rightarrow [d[]_X, a_X] \tag{2}$$

Thus, for two independent frames, as f_i and f_j, the formulation can be realized as,

$$\phi\{ f_i \} \Rightarrow [d[]_i, a_i] \tag{3}$$

$$\phi\{ f_j \} \Rightarrow [d[]_j, a_j] \tag{4}$$

The detection of the unstable video can be detected if the following condition appears in any two consecutive frames as,

$$a_i \neq a_j \tag{5}$$

During such situations, the baseline method recommends calculating the mean angle, a_{Mean} of the total video and further change the angle of the unstable frames with the mean angle. This can be realized as,

$$a_{Mean} \Leftarrow mean\{\sum_{i=1}^{n} a_i\} \tag{6}$$

And further replace with the existing angles, as,

$$a_i \leftarrow a_{Mean} \tag{7}$$

$$a_j \leftarrow a_{Mean} \tag{8}$$

Regardless to mention, the base line method has seen many improvements in recent times. The improvements are discussed in the further section of this work.

III. Foundational Method for Noise Reduction

Secondly, in this section of the work, the baseline method for noise reduction is analyzed. The noise reduction is yet another highly important task before processing the video data for information extraction. Any noisy video data can wrongly influence the decision-making tasks during the processing of the video data.

The baseline method for noise reduction or removal process is furnished here. After realizing the Eq. 1, assuming that, λ is an arbitrary function to extract the size of the pixel information, s_X and list of objects with details, obj_X[] from each frame. This can be formulated as,

$$\lambda\{ f_X \} \Rightarrow [s_X, obj_X[]] \tag{9}$$

The baseline method demonstrates that, the extraction of similar frames, F[], with similar objects is the initial step towards the noise reduction as,

$$F[] \Leftarrow \prod_{obj_X[]=V[i].obj_i[]} V[i] \tag{10}$$

Now assuming that two frames, fi and fj are part of F[] set and contain similar number of objects. Hence, the information size, s_i and s_j, must be same. If the information sizes are different, then it is natural to realize that the frames contain noise as per the baseline method. As,

$$\begin{aligned} &iff\ s_i \neq s_j \\ &then, s_i \,\&\, s_j \Rightarrow Noise \end{aligned} \tag{11}$$

Further, the baseline method indicates to replace the pixel information of the noisy frames with the mean information, S_{Mean} from the similar frames. As,

$$s_i \Leftarrow Mean\{ \sum_{k=1}^{Length\{F[]\}} F[k]\} \tag{12}$$

$$s_j \Leftarrow Mean\{ \sum_{k=1}^{Length\{F[]\}} F[k]\} \tag{13}$$

Regardless to mention, that the base line method has seen many improvements in recent times. The improvements are discussed in the further section of this work.

IV. Foundational Method for Key Frame Extraction

Thirdly, in this section of the work, the baseline method for the key frame extraction is furnished. The video information processing is a highly time complex process in general and to reduce the time complexity of the processing algorithms, the researchers have adapted to method to reduce the size of the video data without losing the critical information from the data. This process is identified as key frame extraction process.

The baseline method indicates, that the total video data must be broken into set of frames, F[], based on a given time threshold as,

$$F[] = \frac{d}{dt} V[] \tag{14}$$

Further, the key frame extraction process, as per the baseline method, is very simple. The consecutive frames contain different information must be considered as key frames, KF[]. This can be formulated as,

$$KF[] \Leftarrow \prod_{F[i] \neq F[i+1]} F[] \tag{15}$$

The terminating condition for such methods is that the length of the extracted key frame set must be less than the total video data as,

$$Length\{KF[]\} \square \ Length\{V[]\} \tag{16}$$

Regardless to mention, that the base line method has seen many improvements in recent times. The improvements are discussed in the further section of this work.

V. Recent Research Reviews

After realizing the baseline methods, in this section of the work, the recent research improvements over the baseline methods are discussed.

To address the problems of miss-election and misselection due to poor video key frame detection, Z. Wang et al. [5] propose an algorithm based on motion vectors to locate the crucial frames in a video. Using the sum of the entropy of the difference between adjacent frames and the entropy of the image in two dimensions, we can quantify frames. Second, the lens boundary can be obtained with the help of statistical tools. ViBe is an algorithm that can recognize the object in the foreground of a video and extract transformation features that are not affected by the size of the image. The motion vector is found by first segmenting two consecutive frames, and then matching those segments block by block. The video's level of motion is reflected in the magnitude of the motion vector, allowing us to identify "active and inactive motion" regions. Video frames are compared using a predetermined formula to determine which regions of frames are most similar, and then video frames are extracted from those regions. Video findings with rich motion information are improved by the proposed detection algorithm, as demonstrated by experimental results. The outcomes of VR key frame extraction are also analyzed in this article.

Determining whether or not a video frame was intentionally deleted or altered, is a crucial part of video forensics. Existing methods are unable to process videos with varying levels of motion. Interfering frames are ignored by these methods. C. Feng et al. [6] aim to develop an interference-free, motion-adaptive forensic technique. Researchers analyze statistically the frames that interfere with one another, like moved I-frames to pinpoint the frames that should be removed (FDPs). The adaptability of the fluctuation feature to various levels of motion is enhanced by the elimination of intra-predictions. The improvement is quantified with the aid of moving window detectors. We propose a post-processing technique to eliminate jarring transitions in brightness, focus, and frame rate. When applied to videos with varying motion intensities and interfering frames, the algorithm has a true positive rate of 90% and a false alarm rate of 0.3%. The proposed technique has potential applications in video forensics.

CRC mortality rates can be lowered through early diagnosis. Polyps are the first stage of cancer. Disease diagnosis and comprehension require analysis of the most crucial endoscopy frames. Sasmal et al. [7] use deep learning to select laparoscopic video key-frames. This technique employs transfer learning due to the scarcity of high-quality polyp depth maps. During an endoscopy, many images are taken. Discarding frames from an endoscopic video that are of poor quality or have no diagnostic value is necessary for making a clinical diagnosis. Key-frame selection is suggested using polyp depth information. This method intelligently chooses key-frames by analyzing the significance of edges, punctuation, and other image features. So that the surgeon can separate the polyp from the mucosa layer, a real-time 3D image of the polyp's surface is provided. Polyps can be more easily pinpointed with the aid of depth maps.

Recognizing anomalies in video footage, which is necessary for uses such as surveillance, is not easy. The state-of-the-art methods for detecting video anomalies typically rely on deep rebuilding models, but their results fall short because there isn't enough of a difference between the reconstruction errors for normal and anomalous video frames to fully optimize the models. Anomaly detection using frame predictions shows promise. Unsupervised content anomaly detection was proposed by X. Wang et al. [8] using frame prediction. To deal with semantically informative objects and regions of varying scales and to capture spatial-temporal dependencies in everyday videos, the proposed method employs a "multipath ConvGRU-based frame prediction" network. In order to lessen the impact of ambient noise, training lowers tolerance for it. The proposed method achieves better results than the current state-of-the-art approaches.

To achieve a higher frame rate in videos, interpolation is used to synthesize new frames between the existing ones. Existing techniques rely on pairing adjacent frames to create intermediate frames, but they struggle with issues like high-velocity motion, occlusion, and blurring. In order to make the most of the spatial and temporal data at their disposal, H. Zhang et al. [9] proposed "a multi-frame pyramid refinement network". There are three technical advancements that would benefit the proposed network. The first step is the proposal of a coarse-to-fine framework for improving optical flows across multiple frames. Estimates can be made for both large-scale motion and occlusion. Second, spatial and temporal context is unearthed, and texture is restored. Third, we use perceptual loss in multiple steps to keep the finer details of the intermediate frames intact. For interpolation between multiple frames, this technique can be used. Overall, 80K frame groups are used in the training process. The method has been shown to outperform state-of-the-art methods and handle challenging cases on multiple independent datasets.

H. Bhuyan et al. [10] extract the key frames from the dance video and the motion frames from the video itself. New,

easy, and efficient, the localization method is suggested. The basic structure of KFs varies depending on the dance style and the dancer. It's not easy to establish a universal threshold for movement that applies to all dancers and performances. The threshold was previously determined using iterative methods.

W. Zhang et al. [11] investigated spatio-temporal video super-resolution. The researchers first propose a cross-frame transformer-based network for end-to-end spatio-temporal video super-resolution. To reconstruct high resolution and frame rate results from coarse to fine, the researchers propose "a multi-level residual reconstruction module" that makes use of the maximum similarity and similarity coefficient matrices produced by the cross-frame transformer. Compared to the standard two-stage network, this method has fewer training parameters while providing improved performance.

The process of "video object extraction (VOE)" identifies and isolates region of interest from a video. To solve the issues of imprecise foreground object extraction and superfluous small-scale motion interference, Y. Guo et al. [12] proposed a novel VOE approach based on "spatiotemporal consistency saliency" detection. The proposed method's main innovation is comprised of three parts: first, the spatiotemporal gradient field (SGF) is built a tempor Experiments on the public video saliency data sets ViSal and SegtrackV2 demonstrate efficiently and accurately identify the salient object in a video sequence.

Semantic sections of a video are extracted using video segmentation, with each segment corresponding to a user-defined concept. The goal of the user has not been taken into account in previous research on video segmentation. With dimension reduction and temporal clustering, X. Peng et al. [13] present a "two-stage user-guided video segmentation" framework. During dimension reduction, coarse-grained features are extracted using ImageNet. During the temporal clustering phase dimensionally, reduced frames is used to segment videos on the time domain based on the user's intended viewing path. To better understand videos, users can employ hierarchical clustering to divide them into smaller, more manageable chunks.

To facilitate the retrieval, it takes advantage of the fact that certain video-related image features tend to exhibit temporal correlations. Key frame identification and key area localization are made possible through the use of video imprint representation. The video imprint tensor is created in the framework developed by Z. Gao et al. [14] by removing redundancy across frames with a specialized feature alignment module. The proposed reasoning network employs a memory-inspired attention mechanism. The latent structure of the reasoning network identifies key frames from the video imprint that can be used to reconstruct the sequence of events. Event retrieval is enhanced over the state-of-the-art techniques thanks to the "compact video representation aggregated from the video" imprint.

Observing moving objects is made possible thanks to video satellite's ability to produce rich actionable information. Time resolution is more important than spatial clarity in video satellite images. The quality of video satellite images is greatly enhanced by super-resolution. "Video satellite image SR reconstruction" was proposed by Z. He et al. [15]. Oftenest calculates the LR optical flow from multiple image frames. Next, an upnet is constructed to enhance the resolution of the input frame and the LR optical flow. Motion compensation is executed in accordance with HR optical flows. Through the use of the HR cube's compensation, the ARLnet produces SR results.

Intelligent service robots with video analysis capabilities are used for complex computing in the cloud. To monitoring human activity, intelligent service robots' film in a continuous loop. "Action classification, recognition, abnormal event detection, and crowd emotion sensing" all require action extraction from unrestricted continuous video. With three components—spatial location estimation, temporal action path searching, and spatial-temporal action compensation. H. Guo et al. [16] proposed a novel approach for action extraction in unconstrained continuous video. It is possible to pinpoint one's location based on a person's outward appearance and their motion. Then, considering missed sightings and false alarms, the results of the spatial action proposal are used to formulate the space - time action trail scanning as an optimal probability model. "The researchers advocate for and demonstrate the convergence of the "Markov Chain Monte Carlo" algorithm.

Because of developments in video and image processing, video tampering forensics is extremely difficult to conduct. Object reduction video forgery requires the use of passive forensics techniques, which are crucial in the courtroom. The "spatial rich model (SRM) and 3D convolution (C3D)" were the basis for the proposal of a spatiotemporal trident network by Q. Yang et al. [17], which is used to extract tampering traces from video. It can enhance the detection and identification of tampered regions, and it has three distinct branches. The "spatiotemporal trident network" served as inspiration for the development of a time-based detector and a space-based locator for locating instances of video manipulation. The temporal detector used 3D convolutional neural networks (CNNs) with three different types of encoders and decoders. C3D-ResNet12 was developed as the spatial locator's encoder. The loss functions of both algorithms were optimized by the researchers.

The future of the space information network will rely heavily on satellite video because of the dynamic information it provides on large spatial and temporal scales. This study employs a two-stream approach to extract EOI from satellite video scenes. Still images, such as scenes, are captured in each frame of a satellite video, while the order of these frames is what establishes motion. In light of these details, we propose a brand new two-stream EOI detection framework. One stream uses AlexNet to pull static spatial data from satellite videos, while the other uses local trajectories analysis to pull data on motion. Before anything else, the video scene is cut up into spatial-temporal patches where EOI and non-EOI regions are labelled. The next step is to take the 3-D satellite video cubes from the event scene patches and extract the trajectories. Finally, this weak-supervision trajectory classification problem is solved by sparse dictionary learning. The experimental outcomes demonstrate the efficacy of the two-stream method in EOI detection and its potential utility in "satellite video analysis" and comprehension. The approach

taken by Y. Gu et al. [18] is superior to current video analysis models.

CNNs can dehaze individual images, as discovered by W. Ren et al. [19]. Scientists investigate the feasibility of using a network to defog video footage. Dehazing a video can make use of the wealth of data available in adjacent frames. Based on the estimated transmission map, a haze-free scene model is created. Since the semantic information of a scene is a powerful prior for image restoration, the team suggests using "global semantic prior to regularize the transmission" maps, making the estimated maps continuous only between objects and smooth within each object. To train this network, scientists generate synthetic hazy and clear videos. Scientists have demonstrated that the features of this dataset can clear up haze in videos of outdoor scenes.

For supervised video summarization, W. Zhu et al. [20] proposed DSNet. Anchors are used and ignored by DSNet. While the anchor-based approach eliminates pre-defined temporal proposals and predicts importance scores and segment locations, the anchor-free approach in contrast to preexisting supervised video summarization techniques, the interest detection framework makes an initial attempt to exploit temporal consistency. Before extracting their long-range temporal features for location regression and importance prediction, researchers in the anchor-based approach. Segments are assigned as positive or negative to ensure the summary is accurate and comprehensive. The anchor-free method directly predicts the significance of video frames and segments, sidestepping the drawbacks of temporal proposals. This framework is compatible with existing supervised video summarization tools. Scientists evaluate both anchor-based and anchor-free methods on SumMe and TVSum. Both the anchor-based and anchor-free strategies have been verified by experiments.

Object recognition is the mainstay of computer vision. There is hope in the use of correlation filters. Because each target is so small and the target and background are so similar, the "kernel correlation filter (KCF) tracker" has trouble keeping up with moving objects in satellite videos. To better locate objects in satellite footage, B. Du et al. [21] suggested combining the KCF tracker with a three-frame-difference algorithm. For a robust tracker, it is suggested to combine the KCF tracker with the three-frame-difference algorithm. Three satellite videos demonstrate the superiority of the proposed method, as demonstrated by the researchers.

Further, based on the recent research improvements, in the next section of this work, the persistent research problems are furnished.

VI. Problem Formulation – Mathematical Model

After the understanding of the baseline methods and detailed analysis of the recent research outcomes, in this section, the persistent research problems are discussed.

Firstly, the problem of information loss due to the change of angle for video stabilization is discussed.

Continuing from Eq. 3 and 4, assuming that, due to the change of the angle with a_X, the frames f_i and f_j, are expected to change the dimensions from $d_i[]$ and $d_j[]$ to $d_X[]$. Naturally, which are different from the original dimensions. This can be realized as,

$$d_i[H_i, W_i] \neq d_X[H_X, W_X] \tag{17}$$

$$d_j[H_j, W_j] \neq d_X[H_X, W_X] \tag{18}$$

Where H and W are the height and width respectively also, it is natural to realize that, if the new dimensions are less than the original dimensions, then the information loss is non-preventable. As,

$$\begin{aligned} &iff\ H_i > H_X\ Or\ W_i > W_X \\ &Then, \lambda(f_i) > \lambda(f_i') \end{aligned} \tag{19}$$

Where, f' is the modified frame after the motion angle change. Hence, it is natural to observe the information loss due to the angle change as per the traditional existing methods.

Secondly, the problem of undetected object mismatch during the noise reduction is realized.

The second issue with the traditional existing methods are to be realized with a condition that, the shape and number of the objects are same, however the position of the objects are different. This can be realized from Eq. 9. Assuming that, two frame f_i and f_j contain the set of objects in the frame as $obj_i[]$ and $obj_j[]$. As,

$$\lambda\{f_i\} \Rightarrow [s_i, obj_i[]] \tag{20}$$

$$\lambda\{f_j\} \Rightarrow [s_j, obj_j[]] \tag{21}$$

Here, the objects in the respective frames are same and the size of the frames are also same as,

$$s_i = s_j \tag{22}$$

$$obj_i[] = obj_j[] \tag{23}$$

Nonetheless, the positions of the objects are different as,

$$obj_i[k].(H, W) \neq obj_j[k].(H, W) \tag{24}$$

Thus, naturally, the information from the frames is also not equal. As,

$$\varpi\{f_i\} \neq \varpi\{f_j\} \tag{25}$$

Where, ϖ is an arbitrary function to extract knowledge from the video frame. Henceforth, the loss of object unique locations over in the noisy frames will be lost as per the Eq. 13.

Finally, the problem of higher time complexity during the key frame extraction is analyzed. As per the initial assumptions, the total number of frames in the video data is “n” and as per the Eq. 15, the total time complexity, T, can be formulated as,

$$T = n*(n-1) \tag{26}$$

Or,

$$T = n^2 \tag{27}$$

Which implies,

$$T(n) = O(n^2) \tag{28}$$

And, during a situation of $n \to High$, naturally, $T(n) \to Very\ High$. Thus, this problem also must be addressed.

Further, the proposed solutions to these identified research problems are furnished in the next section of this work.

VII. Proposed Solutions

After the detailed analysis of the problems with baseline methods and the recent improvements observed to the baseline methods, in this section of the work, proposed methods are furnished here.

Firstly, the proposed video stabilization method without the information loss is realized. The initial process is to identify the variation of the angles in each frame. Using Eq. 2, extracting all the angles from each frame and build the angle collection, A[], as,

$$A[] = \prod V[].a \tag{29}$$

Further calculate the mean angle, A_{Mean} from the angle collection as,

$$A_{Mean} \Leftarrow Mean\{\sum_{i=1}^{n} A[i]\} \tag{30}$$

Further, build another set of frames, FA[], where two consecutive frames angle is different as,

$$FA[] \Leftarrow \prod_{a_i \neq a_j} F[] \tag{31}$$

Further, apply the mean angle to the FA[] collection frames and check the number of frames, where information loss is observed. As,

$$\begin{aligned} &FA[].a \Leftarrow A_{Mean} \\ &iff\ \ \lambda(FA[i]) > \lambda(FA'[i]) \\ &then, C[] \Leftarrow FA[i] \end{aligned} \tag{32}$$

Where, C[] is collection of frames, where after motion stabilization information is lost. Now, if for the maximum number of frames information loss is observed, then the mean angle must be recalculated. As,

$$\begin{aligned} &iff\ \ length\{C[]\} \approx length(FA[]) \\ &then, A_{Mean} \Leftarrow Max\{FA[].a\} \end{aligned} \tag{33}$$

The proposed method indicates to repeat the process until $length\{C[]\} \to 0$. Henceforth, the loss of information is almost reduced to zero.

Secondly, the proposed noise reduction method with object information persistent and background separation is furnished.

The proposed method indicates to separate the background and foreground information from each frame. Assuming that, X is an arbitrary function to perform the separation task, then the following model can be formulated,

$$\Delta\{f_x\} \Rightarrow \{f_x.FG, f_x.BG\} \tag{34}$$

Further, build the set of frames, K[][], with similar background and similar foreground as,

$$K[][] \Leftarrow \prod_{f_x.FG=F[].FG,\ f_x.BG=F[].BG} F[] \tag{35}$$

The mean value of the information extracted from the K[][] collection for the background, BG_{Mean}, foreground, FG_{Mean}, must be calculate to substitute for the noisy frames. As,

$$BG_{Mean} = Mean\{\sum_{i=1,\ j=1}^{length\{K[][]\},n} \lambda\{K[i][j].BG\}\} \tag{36}$$

$$FG_{Mean} = Mean\{\sum_{i=1,\ j=1}^{length\{K[][]\},n} \lambda\{K[i][j].FG\}\} \tag{37}$$

Once the mean values are calculated, then the pixel value information must be replaced with the mean values. Hence, continuing from Eq. 15.

$$f_x.BG \Leftarrow BG_{Mean} \tag{38}$$

$$f_x.FG \Leftarrow FG_{Mean} \tag{39}$$

Hence, now the frames with similar object positions and similar background information can be denoised without missing the specific objects information.

Finally, the proposed variable threshold based key frame extraction method is realized.

The concept of the variable threshold used here is primarily to indicate that as the video information changes frame by frame, thus the selection of the key frames also must be decided dynamically. Hence this proposed method indicates to calculate the keyframe selection method using moving average of the threshold.

Continuing from the Eq. 9 and Eq. 14, the threshold, TH, for the information frame must be calculated as,

$$TH \Leftarrow \lambda\{f_i\} \tag{40}$$

and, further,

$$KF[] \Leftarrow \prod_{F[i].TH > Mean\{\sum_{i=1}^{n} TH(i)\}} F[] \tag{41}$$

Hence, this proposed key frame extraction process reduces the time complexity to O(n) as the mean value calculation is an iterative process along with the same steps of key frame extraction.

Further, based on the proposed methods, in the next section of this work, the proposed algorithms are furnished along with the proposed framework.

VIII. Proposed Algorithms and Frameworks

After the finalization of the proposed mathematical models in the previous section of this work, in this section, the proposed algorithms based on the mathematical models are presented.

Firstly, the Information Loss Preventive Video Stabilization using Mean Angle Measure (ILP-VS-MAM) Algorithm is discussed.

Algorithm - I: Information Loss Preventive Video Stabilization using Mean Angle Measure (**ILP-VS-MAM**) Algorithm

Input:
Video data as V[]

Output:
Motion Stabilized Video data as V1[]

Process:
Step - 1. Read the total video data as V[]
Step - 2. For every frame in V[] as V[i]
  a. Extract the angle as A[i] using Eq. 2.
  b. If A[i] != A[i-1]
  c. Then, C[j] = V[i]
Step - 3. For every element in V[] as V[k]
  a. Calculate the mean angle as MA from A[]
  b. Apply V[k].A = MA and generate V_Temp[i]
  c. If V_Temp[k] contains less information then V[k] using Eq. 32
  d. Then, remove A[k] and C[k] = 0
  e. Else, V1[k] = V[k]
  f. Stop if Length((C[]) = =0)
Step - 4. Return V1[]

In order to prevent visual quality loss, video stabilization technology minimizes unintentional jitters and shakes of an object capturing equipment without affecting moving subjects or purposeful camera panning. This is especially important for handheld imaging devices because of how much more susceptible they are to vibrations. Unwanted changes in camera position led to unstable image sequences, but controlled movement commonly generate unstable images.

Secondly, the Noise Reduction using Object Separation and Background Normalization (NR-OS-BN) Algorithm is discussed.

Algorithm - II: Noise Reduction using Object Separation and Background Normalization (**NR-OS-BN**) Algorithm

Input:
Motion Stabilized Video data as V1[]

Output:
Noise Reduced Video data as V2[]

Process:
Step - 1. Read the total video data as V1[]
Step - 2. For every frame in V1[] as V1[i]
  a. Extract the Background as BG[i]
  b. Extract the Foreground as FG[i]
Step - 3. Calculate the BG_Mean and FG_Mean from BG[] and FG[]
Step - 4. For each information set in BG[] as BG[k]
  a. If BG[k] == BG[k+1] and FG[k] == FG[i+1]
  b. Then, K[][]= V1[k]
Step - 5. For each element in K[][] as K[j][]
  a. If K[j][].BG > BG_Mean and K[j][].FG > FG_Mean
  b. Then,
    i. K[j][].BG = BG_Mean and K[j][].FG = FG_Mean
    ii. V2[]=K[j][]
Step - 6. Return V2[]

The process of reducing noise from a signal is known as noise reduction. Both audio and picture noise reduction methods exist. Algorithms for noise reduction may slightly skew the signal. As with common-mode rejection ratio, noise rejection refers to a circuit's capacity to separate an undesirable signal component from the desired signal component.

Finally, the Key Frame extraction using Moving Average Dynamic Thresholding (KFE-MA-DT) Algorithm is discussed.

Algorithm - III: Key Frame extraction using Moving Average Dynamic Thresholding (**KFE-MA-DT**) Algorithm

Input:
Noise Reduced Video data as V2[]

Output:
Reduced Key Frame set as KF[]

Process:
Step - 1. Read the total video data as V2[]
Step - 2. For each frame in V2[] as V2[p]
  a. Extract the frame threshould as TH[p]
  b. Calculate the Mean Threshold as MT using Eq. 41
  c. If TH[p] > MT
  d. Then, KF[r] = V2[p]
  e. Else, Discard the frame
Step - 3. Return KF[]

Further, based on the proposed algorithms, the framework to automate the process is furnished [Fig. 1].

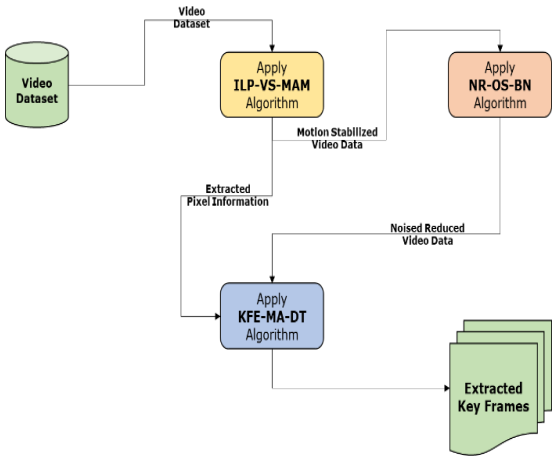

Fig. 1. Proposed Framework.

Furthermore, in the next section of this work, the obtained results are discussed.

IX. Results and Discussions

After the detailed discussions on problems and proposed solutions, in this section of the work, the obtained results are discussed. Firstly, the dataset information is furnished [Table I].

TABLE I. Dataset Information [22]

Attributes	*Number (#)*
Number of Videos	1900
Average number of Frames	7247
Number of Categories	8

The number of samples in the dataset are 1900, however due to representation purposes, only 10 samples are listed. Secondly, the video data stabilization outcomes are furnished [Table II].

TABLE II. Video Motion Stabilization Outcomes

	Before Processing			*After Processing*		
Dataset ID	*Mean Heights*	*Mean Width*	*Dataset ID*	*Mean Heights*	*Mean Width*	*Dataset ID*
1	240	320	1	240	320	1
2	240	320	2	240	320	2
3	240	320	3	240	320	3
4	240	320	4	240	320	4
5	240	320	5	240	320	5
6	240	320	6	240	320	6
7	240	320	7	240	320	7
8	240	320	8	240	320	8
9	240	320	9	240	320	9
10	240	320	1	292	292	2

It is natural to realize that, due to the change of camera angle for each frame, the images are converted to a square image frame and the camera angles are justified. The same results are visualized graphically here [Fig. 2].

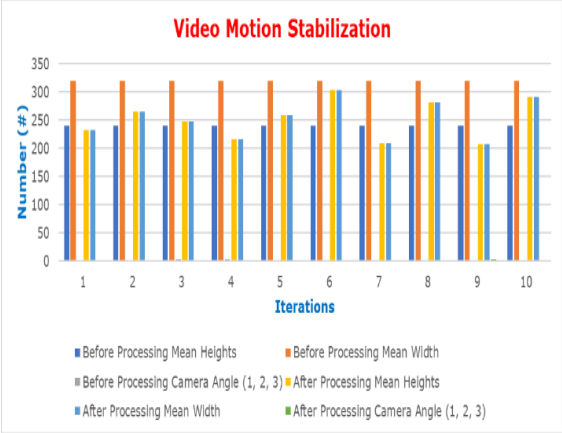

Fig. 2. Video Motion Stabilization Outcomes.

For a single video data, few frame-by-frame analysis are furnished here [Table III].

TABLE III. Single Video Frame-By-Frame Analysis

	Before Processing				*After Processing*			
Dataset ID	*Mean Heights*	*Mean Width*	*Camera Angle (1, 2, 3)*	*FPS*	*Mean Heights*	*Mean Width*	*Camera Angle (1, 2, 3)*	*FPS*
1	240	320	3	10	265	229	2	10
	240	320	3	10	259	269	3	10
	240	320	3	10	212	251	2	10
	240	320	3	10	282	186	1	10
	240	320	3	10	293	173	2	10
	240	320	3	10	276	114	2	10
	240	320	3	10	201	247	2	10
	240	320	3	10	293	235	3	10
	240	320	3	10	131	175	2	10
	240	320	3	10	248	168	2	10

Thirdly, the analysis of the noise reduction process is analyzed [Table IV].

TABLE IV. Video Motion Stabilization Outcomes

Dataset ID	*Mean Initial Noise (dB)*	*Mean Reduced Noise (dB)*	*Time (ns)*
1	5.7684727	2.7311466	1.722
2	4.3735304	1.9569633	7.556
3	5.286506	3.096948	7.395
4	4.7074976	2.7789779	7.110
5	4.768707	2.8522859	3.160
6	4.387303	2.6379135	9.517
7	4.337146	2.4204903	6.397
8	3.9589996	2.2259648	8.053
9	3.7900214	2.1423862	1.512
10	3.7343879	2.0677133	6.911

Thus, it is natural to realize that the noise reduction [Fig. 3] with a minimal time complexity [Fig. 4] is achieved.

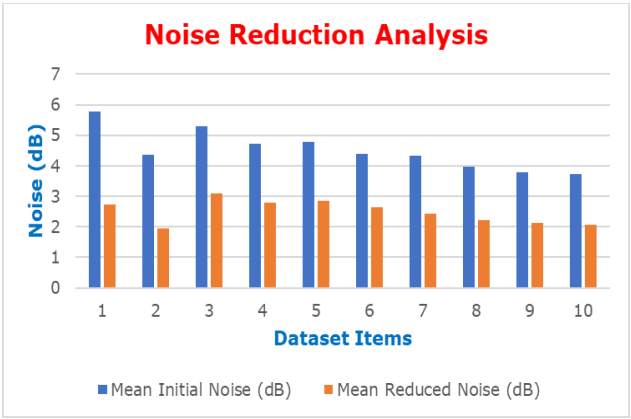

Fig. 3. Noise Reduction Analysis.

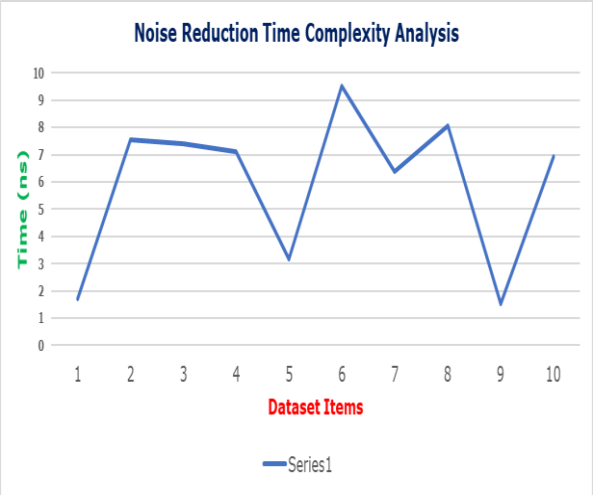

Fig. 4. Noise Reduction Time Complexity Analysis.

Finally, the key frame extraction outcomes are analyzed [Table V].

TABLE V. KEY FRAME EXTRACTION ANALYSIS

Dataset ID	*Initial Number of Frames*	*Number of Key Frames*	*Adaptive Threshold*	*Time (ns)*
1	150	119	90456	8.788
2	61	56	96709	4.229
3	181	83	98426	6.131
4	141	32	99201	7.822
5	131	61	94871	3.244
6	81	16	95059	6.798
7	151	148	95206	5.393
8	71	64	91944	6.577
9	111	61	90952	7.671
10	151	72	90036	2.254

Henceforth, the reduction in terms of key frames [Fig. 5] is clearly visible with finite number of iterations, which again reflects in the minimal time complexity [Fig. 6].

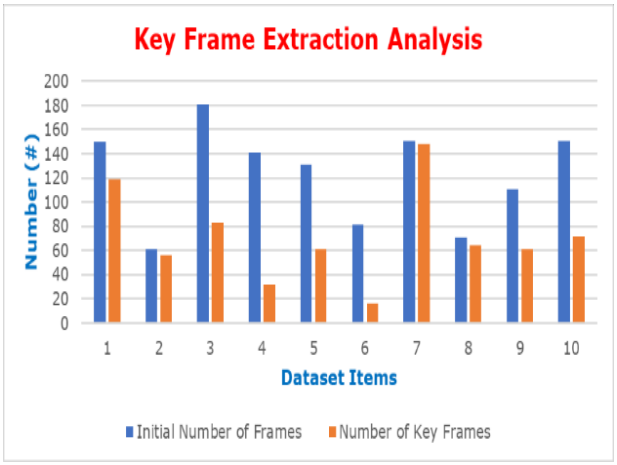

Fig. 5. Key Frame Extraction Analysis.

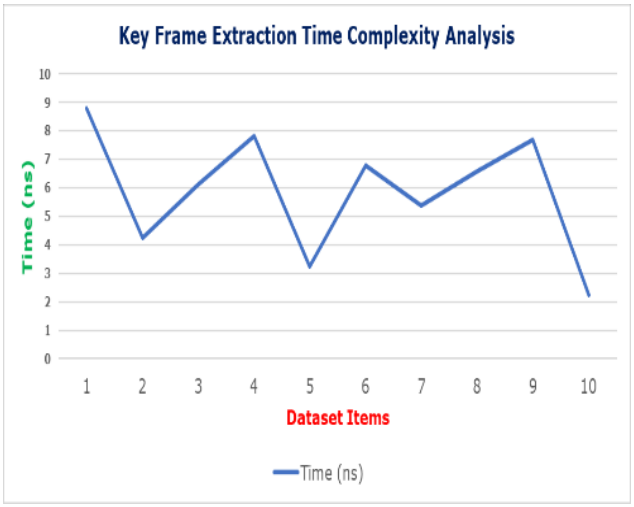

Fig. 6. Key Frame Extraction Time Complexity Analysis.

Henceforth, with the complete confirmation of the achieved results discussed in this section of the work, in the next section the obtained results are compared with other parallel research works.

X. COMPARATIVE ANALYSIS

After the detailed discussion on the proposed solutions, algorithms and the results, this section of the work, is dedicated to summarizing the outcomes and further compare with significant parallel research works [Table VI].

TABLE VI. COMPARATIVE ANALYSIS

Author, Year	*Proposed Method*	*Key Frame Reduction (mean) (%)*	*Model Complexity*	*Time Complexity (mean) (ns)*
Y. Zhang et. al. [2], 2020	Key Frame Extraction	49%	$O(n^2)$	9.58
P. Sasmal et. al. [7], 2021	Key Frame Extraction	51%	$O(n^2)$	7.55
X. Wang et. al.[8], 2022	Key Frame Extraction	63%	$O(n^2)$	7.85
Proposed Method	Motion Stabilization, Noise Reduction, Key Frame Extraction	65%	O(n)	5.89

It is natural to realize that the proposed method has outperformed the parallel research works. Henceforth, in the next section of the work, the research conclusion is presented.

XI. CONCLUSION

Over the past decade, research into video information processing has been one of the most concentrated efforts, with many studies focusing on pre-processing concerns. Researchers are unable to use information retrieval or categorization algorithms due to pre-processing problems such as inconsistent video frame rates or collection angles, noisy data, and enormous file sizes. Due to the popularity of using video as a means of presenting information, addressing these

issues is equally crucial. The purpose of this study is to automate the processes of motion stabilization, noise reduction, and key frame extraction so that these tasks can be completed in less time and with fewer losses of information. This work initially applies the Information Loss Preventive Video Stabilization using Mean Angle Measure (ILP-VS-MAM) Algorithm to stabilize the motion across the frames, secondly applies the Noise Reduction using Object Separation and Background Normalization (NR-OS-BN) Algorithm to reduce the noise from each frame and from the overall video data and finally, applies the Key Frame extraction using Moving Average Dynamic Thresholding (KFE-MA-DT) Algorithm to extract the minimal amount of key frames with reduced time to complete all the processes. The study leads in a 66% reduction in key frame extraction time and a duration of roughly 6 ns for processing all of the video data.

REFERENCES

[1] Y. Li, A. Kanemura, H. Asoh, T. Miyanishi and M. Kawanabe, "Multi-Sensor Integration for Key-Frame Extraction From First-Person Videos," in IEEE Access, vol. 8, pp. 122281-122291, 2020.

[2] Y. Zhang, D. Zhu, H. Bi, G. Zhang and H. Leung, "Scattering Key-Frame Extraction for Comprehensive VideoSAR Summarization: A Spatiotemporal Background Subtraction Perspective," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 7, pp. 4768-4784, July 2020.

[3] C. Huang and H. Wang, "A Novel Key-Frames Selection Framework for Comprehensive Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 577-589, Feb. 2020.

[4] Q. Zhong, Y. Zhang, J. Zhang, K. Shi, Y. Yu and C. Liu, "Key Frame Extraction Algorithm of Motion Video Based on Priori," in IEEE Access, vol. 8, pp. 174424-174436, 2020.

[5] Z. Wang and Y. Zhu, "Video Key Frame Monitoring Algorithm and Virtual Reality Display Based on Motion Vector," in IEEE Access, vol. 8, pp. 159027-159038, 2020.

[6] C. Feng, Z. Xu, S. Jia, W. Zhang and Y. Xu, "Motion-Adaptive Frame Deletion Detection for Digital Video Forensics," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 12, pp. 2543-2554, Dec. 2017.

[7] P. Sasmal, A. Paul, M. K. Bhuyan, Y. Iwahori and K. Kasugai, "Extraction of Key-Frames From Endoscopic Videos by Using Depth Information," in IEEE Access, vol. 9, pp. 153004-153011, 2021.

[8] X. Wang et al., "Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 6, pp. 2301-2312, June 2022.

[9] H. Zhang, R. Wang and Y. Zhao, "Multi-Frame Pyramid Refinement Network for Video Frame Interpolation," in IEEE Access, vol. 7, pp. 130610-130621, 2019.

[10] H. Bhuyan, P. P. Das, J. K. Dash and J. Killi, "An Automated Method for Identification of Key frames in Bharatanatyam Dance Videos," in IEEE Access, vol. 9, pp. 72670-72680, 2021.

[11] W. Zhang, M. Zhou C. Ji, X. Sui and J. Bai, "Cross-Frame Transformer-Based Spatio-Temporal Video Super-Resolution," in IEEE Transactions on Broadcasting, vol. 68, no. 2, pp. 359-369, June 2022.

[12] Y. Guo, Z. Li, Y. Liu, G. Yan and M. Yu, "Video Object Extraction Based on Spatiotemporal Consistency Saliency Detection," in IEEE Access, vol. 6, pp. 35171-35181, 2018.

[13] X. Peng, R. Li, J. Wang and H. Shang, "User-Guided Clustering for Video Segmentation on Coarse-Grained Feature Extraction," in IEEE Access, vol. 7, pp. 149820-149832, 2019.

[14] Z. Gao, L. Wang, N. Jojic, Z. Niu, N. Zheng and G. Hua, "Video Imprint," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 12, pp. 3086-3099, 1 Dec. 2019.

[15] Z. He, J. Li, L. Liu, D. He and M. Xiao, "Multiframe Video Satellite Image Super-Resolution via Attention-Based Residual Learning," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-15, 2022, Art no. 5605015.

[16] H. Guo, X. Wu and N. Li, "Action Extraction in Continuous Unconstrained Video for Cloud-Based Intelligent Service Robot," in IEEE Access, vol. 6, pp. 33460-33471, 2018.

[17] Q. Yang, D. Yu, Z. Zhang, Y. Yao and L. Chen, "Spatiotemporal Trident Networks: Detection and Localization of Object Removal Tampering in Video Passive Forensics," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 4131-4144, Oct. 2021.

[18] Y. Gu, T. Wang, X. Jin and G. Gao, "Detection of Event of Interest for Satellite Video Understanding," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 11, pp. 7860-7871, Nov. 2020.

[19] W. Ren et al., "Deep Video Dehazing With Semantic Segmentation," in IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1895-1908, April 2019.

[20] W. Zhu, J. Lu, J. Li and J. Zhou, "DSNet: A Flexible Detect-to-Summarize Network for Video Summarization," in IEEE Transactions on Image Processing, vol. 30, pp. 948-962, 2021.

[21] B. Du, Y. Sun, S. Cai, C. Wu and Q. Du, "Object Tracking in Satellite Videos by Fusing the Kernel Correlation Filter and the Three-Frame-Difference Algorithm," in IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 2, pp. 168-172, Feb. 2018.

[22] Dataset: M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, June 2016.