Date of publication DEC-31, 2023, date of current version MMM-10, 2023.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.22.4.3352

Investigating Methods of Searching for Key Frames in Video Flow with the Use of Neural Networks for Search Systems

NATALYA SHAKHOVSKA, NATALIA MELNYKOVA, PETRO POBEREIKO, MARYANA ZAKHARCHUK
Lviv Polytechnic National University, Lviv, Ukraine

Corresponding author: Natalya Shakhovska (e-mail: natalya233@gmail.com).

ABSTRACT Various methods of video content data analysis are presented, compared, and evaluated in this paper. Due to the analysis, the most effective strategies for video data processing involve searching for key frames within the video stream. The examined methods are categorized into consistent comparison, global comparison based on clustering, and event/object-based methodologies. Key techniques such as sequence search, classification, frame decoding, and anomaly detection are singled out as particularly valuable for comparison and matching tasks. The research further reveals that artificial intelligence and machine learning-driven methods reign supreme in this domain, with deep learning approaches outperforming traditional techniques. The employment of convolutional neural networks and attention mechanisms to capture the temporal intricacies across variable scopes is especially noteworthy. Additionally, leveraging the Actor-Critic model within a Generative Adversarial Network framework has shown encouraging outcomes. A significant highlight of the study is the proposed approach which incorporates modified Independent Recurrent Neural Networks (IndRNN) complemented by an attention mechanism. The enhancement using mathematical tools, notably the standard deviation, for key frame detection, exemplifies the potential of integrating analytical instruments to refine the system's precision. Such advancements, as presented in this research, pave the way for substantial enhancements in information systems tailored for video content analysis and source identification.

KEYWORDS key frames; neural networks; unsupervised learning; SIFT; CNN; IndRNN; Leaky ReLU.

I. ARTICLE FORMATING

With the rapid development of the Internet and modern technologies, image and video networks have become one of the most effective ways of transmitting data. This is due to the difference in processing visual imagery and complex textual content by human brain. While it is true that reading also involves vision, cognitive processing differs significantly. For example, an average person perceives approximately 80% of information through visual imagery, which is often more immediate and intuitive, compared to the 20% from the more analytical and sequential process of reading text. Therefore, finding important information in large texts can take considerable time. Video content leverages this preference for visual learning to quickly and effectively convey information that might be less apparent or more time-consuming to grasp through text content alone. That is why a lot of video content on a wide variety of topics is nowadays distributed on the Internet, and it can be viewed at the same time by millions of people in different places worldwide. Besides, we can see that on various streaming services, in addition to the original video, there are usually its separate fragments, which in many cases are distorted or damaged. Therefore, due to the above, a research devoted to the improvement of known and the development of new methods of video content analysis for the search of the original video, but not its distorted or damaged copies and individual fragments, is extremely relevant. The need to solve this problem has prompted many scientists to develop various approaches and algorithms for content analysis, classification, retrieval, and generalization of visual content. Many of them managed not only to build these algorithms, but they also succeeded in discovering the basic principles on which they should be based. In particular, they showed that artificial intelligence and machine training methods were powerful tools for improving known and building new algorithms. It has been revealed that the development of these methods contributes to the rapid development of information systems, with the help of which it is possible to successfully analyze video content and recognize

the original. They have found that it is expedient to carry out video stream analysis based on the results of comparison and search for matches in the sequence of frames (fragments). Objects on the stage or a color palette can serve as frame similarity. It has been established that the basis of modern information systems for the analysis of video content is mostly the use of the so-called algorithms for determining key frames. And based upon the analysis of the methods of their practical application, it has been proved that they make it possible to form a representative sample of a concise review, which provides the most accurate display of video content. Besides, it has been revealed that effective video generalization with highlighting key frames greatly facilitates viewing and navigation of large collections of videos on the Internet, which significantly increases viewer's engagement and content consumption, as well as serves a key step in fragment search systems.

Thus, for solving the given problem, an analytical review of approaches and methods of analyzing videos and fragments, which are based on the methods of artificial intelligence and machine training, as well as the search for existing concepts of visual content generalization, is extremely important. In order to conduct such a review, it is necessary to analyze the following methods: search for sequences (detection of objects or certain actions on frames), classification (determination of the content of frames and their distribution into certain categories), decoding of frames (description of the characteristics of a specific image), detection of anomalies (search for objects and symbols that are unique properties of the fragment relative to others).

The main contribution of this paper is as follows:

1) The modified Independent Recurrent Neural Networks (IndRNN) with Leaky Rectified Linear Unit (LeakyReLU) activation function are proposed. Based on that modification our model can assign different weights to different frames, effectively focusing on more "attention-worthy" frames, potentially key frames and achieve in 1.13 times better result in testing dataset than Generative Adversarial Networks (GAN) and Convolutional Neural Networks (CNN).
2) The innovative approach is employed to further enhance the performance of the proposed system. The essence of this approach lies in analyzing standard deviations of feature vectors within a specific frame window. This allows for a significant reduction in noise and an increase in the accuracy of key frame selection. Since some segments of visual data are characterized by high variability, including changes in events and objects, to ensure the required precision of the results, the weight values in the neural network have been increased and prioritized.

II. RELATED WORKS

To study the methods of searching for key frames in a video stream, we shall analyze the functions of analysis of forms, colors, and the optical stream.

In general, there are quite a lot of methods of searching for frames in a video stream. All of them have an important practical significance. In some cases, it is expedient to apply these methods, and in other cases other methods. Application of one or another method of visual data processing is determined by the structure of these data. Therefore, in order to simplify their analysis, we shall theoretically divide them into the following categories: sequential comparison, global comparison, based on clustering, and those using events or objects.

Methods of the sequential comparison category are mainly applied to solve the problems of determining the similarity between frames of a video stream. Their essence is to compare each subsequent frame with the previous one. Quite often, this comparison uses a color histogram, since the difference in the colors of the frames in the video fragment that represent a certain event is mostly insignificant (minimal). However, the methods of this category have some drawbacks. The main of them is the large amount of time spent on processing frames of the video stream and a large error in the case of processing data with noise.

For a brief review of modern methods that are the closest to solving the problems that the visual data generalization system must perform, we shall divide them into two types: conventional methods and methods applying machine training.

The conventional methods are based on a specific objective function and instructions that do not change over time. The vast majority of these functions use mainly pipelined segmentation. Usually, such approaches highlight the SIFT characteristics and an optical stream. Key points and local characteristics of frames are highlighted by means of SIFT descriptors. For example, in the work "Robust voting algorithm based on labels behavior for video copy detection" [1], frame reference points are selected (by means of the Harris detector) and their positions are tracked throughout the video. After that, an infinitely large number of trajectories of these points are formed. Fuzzy search algorithms are used to find similarities. This method significantly simplifies the localization of vague duplicate fragments and enables generalization for the video stream. However, it is resource-intensive for highlighting key points in images. And the fact that the trajectories of the points are sensitive to the movement of the camera makes the algorithm optimal only in cases of searching for exact copies of a video. Another no less important type of conventional approaches are methods based on clustering, the characteristic feature of which is that the number of clusters, as a rule, must be specified a priori. Important in terms of improvement and development of such new methods were given in paper prepared by Tang [2], where it was proposed for the first time to use entropy and density of image grouping for recognition of hand gestures to determine the key frame in the video stream. No less valuable are the studies by Vazquez, who in his paper [3] suggested finding key frames using a method based upon spectral clustering. The uniqueness of this method is that for its practical implementation it is not necessary to determine the degree of similarity with the selection of common features for two images. Also relevant is the paper of Wang [4], in which a method was synthesized, the idea of which is to calculate the similarity matrix and determine clusters based on it with further search and extraction of key frames. A characteristic feature of this method is that it eliminates the restriction of the selection of one frame per cluster. The number of frames selected from one class depends on the length and complexity of the content of the scenes. Compared to key frame search methods based on classification, this method is much simpler from a computational point of view. Its main drawback is that the conclusion about the significance of the frames is made based on the premise that the camera focuses longer on significant scenes. When key frames are selected from a long sequence of frames in a cluster, the middle frame of each sequence is

considered significant, which reminds of the earliest approaches to finding key frames.

In order to eliminate limitations and improve the effectiveness of conventional methods, we shall review works that are devoted to the synthesis of models using deep machine training (multilayer neural networks). One of such studies is the work by Yang [5], which presents a bidirectional short-term memory Bi-LSTM, which was used by configured Graph Attention Networks (GAN) to highlight reference frames from video. This network allowed the authors to transform the visual functions of the image into higher-level functions using the mechanism of Contextual Features-based Transformation (CFT).

The next important scientific work is the work by Mahasseni [6], where the generative competitive network (GCN) was used for the first time. The result of the study was the development of a network of summaries for training to minimize the distance between educational videos and the distribution of their generalizations. The model consisted of a LSTM auto-encoder as an adder and another auto-encoder as a discriminator. Thus, the auto-encoder summarizer was trained to mislead the discriminator, causing the summarizer to produce better summaries. Such a model demonstrated high results during experiments.

The work by Jian [7] describes the peculiarities of using local functions, in particular, functions of scale-invariant feature transformation (SIFT) and functions obtained from a convolutional neural network (CNN). Besides, an integrated function matching scheme was proposed, integrating the matching of SIFT functions and CNN functions between images in order to detect partial copies of images. In this scheme, the authors implemented SIFT feature mapping based on the visual word model to detect potential duplicate region pairs between images, and then mapped the CNN features of these regions extracted from the convolutional layer of the CNN network to calculate the image similarity. Also, a deep generalization network using reinforcement training was used in the work by Zhou et al. (2018) [8]. Th e authors presented a RL-based DQSN approach for video generalization. Based on an unsupervised and supervised training experiment, their objective function enables semantic analysis using only easy-to-obtain labels at the video level.

Some more important studies in this area of knowledge are based on the results of the analysis of spatio-temporal relationships between parts of the video and the use of modern types of neural networks, in particular convolutional networks. Important in this area of knowledge is paper [9], the authors of which, using convolutional neural networks LSTM, synthesized the architectural model of the decoder and encoder. This model makes it possible to simulate the spatial-and-temporal relationship between video fragments. The algorithm proposed on its basis effectively generates a visual variety of key frames and the mechanisms for finding them.

In 2019, Yuan published his paper [10], in which he developed a method that allows training the system and creating a new representation using a fusion strategy. To evaluate a series of consecutive frames, the authors used loss functions. In the same year, Elfeki [11] and the authors of work [12] built a model, in which they successfully combined CNN and Gated Recurrent Units (one of the RNN types). A characteristic feature of this model is the generation of vectors to evaluate the level of activity and the importance of each frame in the video stream.

Generally, the authors of the studies consider that methods for summarizing and finding key frames in a video are most effective only in cases of using machine training.

III. METHODS

A comparative analysis of literary sources has revealed that today there are a number of methods used in modern research for video generalization and which, in our opinion, are the most productive in the case of their application in fragment search systems.

Taking into account the current state of the subject area, we are convinced that the methods that use machine unsupervised training are ideal for video fragment search systems. They make it possible to develop models that during data processing are independent of user intervention.

Most of the modern approaches typically involve the following general steps to extract key frames:

1. Data pre-processing. The video data needs to be pre-processed before feeding into the neural network. This may involve resizing the frames to a consistent resolution, normalizing pixel values, and organizing the frames into a suitable input format for the network.
2. Designing the neural network architecture. It is needed to choose or design neural network architecture suitable for the task of key frame extraction. This can include various types of convolutional neural networks (CNNs) or recurrent neural networks (RNNs), depending on the specific requirements and characteristics of the video data.
3. Training the neural network. For the neural network's training, a sizable labeled dataset of videos with annotated key frames is required. To train the network the patterns and characteristics that separate key frames from non-key frames, methods like backpropagation and gradient descent are used.
4. Key frame prediction. Once the neural network has been trained, it may be used to anticipate the critical frames in videos that have not yet been viewed. The network analyzes the video frame by frame and generates a probability or score for each frame that indicates how likely it is to be a critical frame.
5. Post-processing and thresholding. The projected probability or scores can be thresholds to identify the key frames. Applying a threshold value will exclude frames with low probability. The choice of key frames can also be improved by using post-processing methods like non-maximum suppression or temporal coherence.

There are three basic categories into which these methods may be divided: shot-based, sampling-based, and clustering-based strategies. Shot-based – in this method, the shot boundary/transition is initially detected using an effective Sentence Boundary Disambiguation (SBD) method. This method is a technique used in natural language processing (NLP) to accurately identify the boundaries of sentences in a text. This is a crucial task because effective functioning of many NLP applications, like machine translation, text summarization, and sentiment analysis, depend on correctly segmented sentences. SBD in the context of key frame extraction plays a role in this process in scenarios where textual content is involved, such as videos with subtitles or embedded textual information. The key frame extraction is then carried out once the video frames have been divided into multiple shots. Different key frame selection methods have been

described in various journals. The first and last frames of the prospective shot are traditionally chosen as the crucial frames [13]. These culled key frames are the shots' representative frames, which result in a more simplified synopsis of the original video. Sampling-based – video content is not prioritized, representative frames are selected by equally or randomly sampling the video frames from the original video. The idea behind this method is to choose every k*th* frame from the source video. The length of the video determines this value of k. A typical range for a video summary is 5% to 15% of the entire video. Every 20th frame is chosen as the key frame in the case of 5% summary, whereas every 7th frame is chosen as the key frame in the case of 15% summarization [14]. These key frames are just part of the original video material, and it is possible that some unnecessary frames can also be included. Clustering-based strategies – unsupervised learning techniques such as clustering, grouping clusters of related data points together. With this technique, video file frames with comparable visual contents are divided into various numbers of clusters. The frame that is retrieved as the key frame from each cluster is the one that is the closest to the candidate cluster's center. The characteristics of the frames, such as color histograms, texture, saliency maps, and motion, are what define how similar they are to one another. The fundamental problem with the clustering-based approach is that, before completing the clustering process, it might be challenging to count the number of clusters in a given video clip.

Most approaches, which are based upon unsupervised training, use the rule that a representative sample should help the user or other system components make inferences about the original video content. In this context, methods usually use GAN. This generalization consists of a key frame selector (importance is evaluated) and a generator (it generates a report). The peculiarity of the application of these approaches is that the training takes place by reconstructing the video based on the summary [15]. Let us consider the concept of this approach during training, the essence of which is clearly shown in Fig. 1. Usually, the adder consists of a key frame selector, which evaluates their importance, and a generator, the main task of which is video reconstruction. The discriminator is taught video reconstruction together with the original data as the input data (which results in providing a similarity score). The training process is as follows: the adder tries to fool the discriminator while it tries to learn to find the difference between the summarized report (key frames) and the original video. The result of the training is the state of the discriminator when it cannot find the difference (the classification error is approximately the same for both the reconstructed and the original video).

For the concept demonstrated above, the video reconstruction process based on the summary is described. In order for the model to perform a reverse process, which can later be used as a component in search systems, modern research offers an extended approach. In this case, a pair of discriminators is used. Then the frame selector (bidirectional LSTM) finds the key frames by modelling the temporal dependence between the frames. Further, on the basis of the results of the selector, evaluation is made, which consists of two GANs [16]. The first one is used for training in relation to video reconstruction based on generalization, and the second one trains to perform actions in reverse – from the original to the key frames. Now, let us consider the functioning of the GAN second module in more details. Today, several versions of the training process are offered. One of them is the use of the "actor-critic" model, which considers the task of selecting reference frames as the task of generating sequences. The "actor" and "critic" [17] are in a state of constant exchange of results, let us call it a competition. The training strategy allows the critic to train the value function and the actor to effectively train about the politics of choosing key fragments. This helps us to adjust the selection of the correct values for the model parameters. High efficiency is also demonstrated by GANs, which are based on self-control [18]. In this case, the generator attempts to predict a frame-level importance score at the frame level for each frame and creates weighted frame features based on temporal representations of the raw frame features. The raw frame features and the weighted frame features are then treated as real and fake input data for the discriminator to perform a comparison. It should be considered that to capture time dependences at a long distance throughout the entire interval of the video sequence BiLSTM is used [19]. The drawback of these concepts is still the instability of the training process [20], although in recent studies an attempt has been made to constantly expand and improve GAN models in order to minimize the limitations of the frame evaluation criteria.

The general concept of reinforcement training is based on the function of "rewards". It is divided into several phases. Based on input data in the form of a sequence of frames, a report is generated by predicting an importance score at the level of each frame. The generated report is sent to the next module, which is responsible for quantifying the existing (pre-selected) characteristics using manually created "reward" functions. Then, the calculated scores are combined to form a total value of the "reward" that the adder uses for training. One of the approaches is to train the adder in such a way that it creates diverse and representative samples of key frames using a reward for diversity. This award (cores or coefficient) measures the difference between the selected key frames, and the award for representativeness calculates the distance (that expresses the visual similarity) between the selected frames from the remaining video frames. Also, to solve the problems of LSTM with respect to the vanishing and exploding gradients, independent recurrent neural networks (IndRNN) [21] are used based on the activation function – Leaky ReLU (Leaky Rectified Linear Unit) [22].

The general concept of analysis of motion of key visual objects is based on an auto-encoder, where steps are performed during the analysis, the purpose of which is to find reference objects and their motion trajectories. After having performed these steps, segmented motion video fragments are created for each object. At the next phase, the online model of the auto-encoder of the object (Stacked Sparse LSTM Auto-Encoder) [23] is used to remember the previous states of the object's motion by way of constantly updating the adapted auto-encoder network. The last step is responsible for the fragment reconstruction.

After building the neural model, it is also worth considering the optimization of the network and its functions. One way to optimize is to use a technique called transfer learning. This involves training on a large dataset of images or videos and then using the learned features to analyze a new dataset. This can be much more efficient than training from scratch on a new data set because neural networks can learn useful features faster.

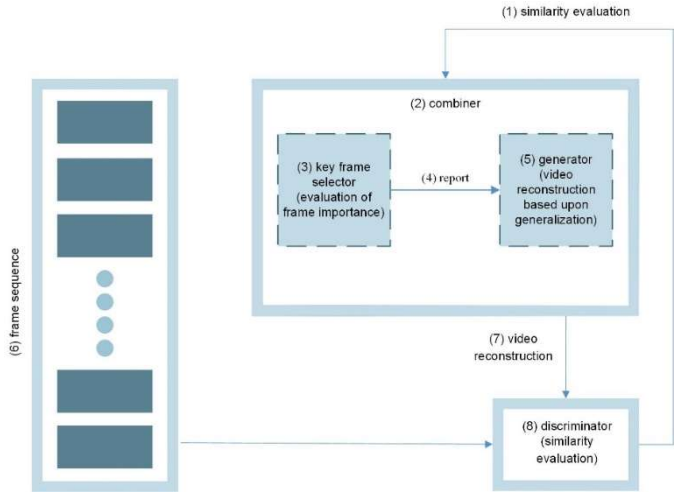

Figure 1. High-level representation of the unsupervised training model for generalization analysis

One more important concept in research is reduction that can be used for optimization. This is the process of removing neurons from the network that do not contribute much to the final output. This can help reduce the computational cost of running a neural network (NN), and can also make the NN more interpretable, making it easier to understand how the network makes its predictions. Another technique that can be used to optimize NNs for video stream analysis is data augmentation. Data augmentation involves creating new training examples by applying various transformations to existing data. This can include techniques such as cropping, flipping, rotating, and adding noise to images. By training a NN on augmented data, it can become more robust to variations in the input data, which can be useful for analyzing a video stream where there can be a lot of variation in lighting, camera angle, and other factors.

IV. RESULTS AND DISCUSSION

Based on the analysis of the literature sources provided above, it has been found that, to date, the most effective types of neural networks for key frame identification in video streams are CNN, GAN, and IndRNN. Let us compare them and discuss their efficacy in practical application:

1. Convolutional Neural Networks (CNN): These are widely used for key frame extraction and visual content analysis. CNNs process image data into a hierarchical model, which is crucial for key frame identification. In tests, CNNs demonstrate high accuracy in key frame identification, attributed to their capability to capture visual patterns and spatial dependencies.
2. Generative Adversarial Networks (GAN): These are employed for creating visually representative frames through generator training. Moreover, with the help of a discriminator, they are also applied to distinguish between generated and actual frames. One of the main advantages of GANs is their ability to generate new frames and effectively convey the content of video materials. However, there is a challenge in finding the necessary criteria to strike a balance between ensuring the integrity of key frames and their representative sampling. This challenge might be reduced in many cases. For a complete realization of their potential and its reduction, further tuning and meticulous training may be required.
3. Independent Recurrent Neural Networks (IndRNN) are a subtype of Recurrent Neural Networks (RNN) that allows neurons to operate autonomously. IndRNNs efficiently model sequential data with a notable ability to capture temporal dependencies in video. They can consider the temporal context and dynamics between frames, assisting in key frame identification. However, due to their focus on temporal modeling instead of explicit image analysis, IndRNNs may be less accurate compared to CNNs and GANs in terms of precision [25]. Nonetheless, it is worth noting that this study emphasizes video content analysis, where the temporal sequence and dependence of images are of significance.

For the practical application of CNN, GAN, and IndRNN neural networks, the fundamental activation functions are a key aspect and should possess several crucial properties:

1. Non-linearity: The model should have the ability to formulate non-linear features, utilizing activation functions like Leaky ReLU [27]. Such functions provide the model with the capability to represent more intricate relationships between pixels and objects. Considering that key frames may encompass scenes with varying lighting, background, or objects, such non-linearity is beneficial for their identification.
2. Informational Value: Activation functions in neural networks play a pivotal role in information relay. Certain pixels or features can be suppressed by activation functions, diminishing their significance in determining key frames. However, pixels containing vital information for key frame identification can be activated through specific activation mechanisms.
3. Gradients and Learning Process: Choosing a specific activation function can significantly influence the learning rate and the stability of the model's convergence. Using certain activation functions can help avoid issues related to vanishing or exploding gradients, ensuring a more stable and efficient learning process.

Based on the conducted research, it has been determined that contemporary video content analysis requires refined methods for key frame detection to optimize search processes.

The emphasis on key frames is motivated by their ability to make video content compact. They serve as the primary tools for summarizing its content. In line with the identified needs and challenges, this study has developed a novel approach to architecture that integrates the consistent features of Independent Recurrent Neural Networks (IndRNN) with adaptive functions of the attention mechanism.

Based on the conducted research, it has been established that IndRNNs, due to their structure, can effectively process prolonged sequences, making these networks suitable for analyzing video sequences with temporal dependencies. However, as previously mentioned, standard IndRNNs have certain limitations, especially in the context of non-linear sequences typical of dynamic videos. To overcome these limitations, modifications were made to the connections in the neural network, and they were integrated with the LeakyReLU activation function. LeakyReLU provides a slight gradient even for negative input values, addressing the vanishing gradient problem, making it ideal for our modified IndRNN. The attention mechanism, recognized for its ability to focus on key features, became a logical addition to our system. In video materials, not all frames are equally important. By using Attention, our model can assign different weights to different frames, effectively focusing on more "attention-worthy" frames, potentially key frames.

To further enhance the performance of the proposed system, an innovative approach is employed, the essence of which lies in analyzing standard deviations of feature vectors within a specific frame window. This allows for a significant reduction in noise and an increase in the accuracy of key frame selection. Since some segments of visual data are characterized by high variability, including changes in events and objects, to ensure the required precision of the results, the weight values in the neural network were increased and prioritized.

The main stages of the proposed approach are:

Noise Reduction:

1. Temporal Smoothing: Video sequences might contain temporal deviations caused by minor object movements, lighting changes, and compression artifacts. These slight variations can sometimes be interpreted as significant content changes, even though they are trivial in reality. Such misinterpretations can be avoided by averaging feature vectors, thus minimizing the noise impact.
2. Frequency Filtering: Video signal is transformed into a frequency spectrum, using methods such as the Fast Fourier Transform, to identify and suppress noise frequencies.

Feature Variability Analysis:

1. Standard Deviation Analysis: Determining the variability or content changes in frames based on the calculation of the standard deviation of feature vectors within a frame window. The larger the standard deviations, the more likely there is a scene change or action, which is one of the most crucial criteria for finding a key frame.
2. Scene Transition Graph: Analyzing transitions between frames is a critical step in understanding the dynamics of video content, as it represents key moments of scene change or highlights main image elements. To further study these transitions, we developed a graph structure where each node symbolizes a specific frame, and the edges between nodes represent the degree of changes between adjacent frames. This approach not only allowed us to detect key transitions between scenes but also to quantify their presence and intensity. Consequently, through this analysis, the proposed system acquired the ability to automatically identify key frames, especially those located at nodes with the most pronounced transition variability.

Integrating these stages into the proposed system provides a comprehensive approach to key frame selection and significantly enhances the accuracy metrics of key frame identification. To validate this conclusion, several brief experiments are conducted with publicly available neural network architectures in the study.

For these experiments, the UCF-101 dataset was used, which contains 13,320 clips with a fixed frame rate of 25 and a resolution of 320 x 240 pixels. The data were divided into training, validation, and testing sets, comprising 70% training data, 15% validation data, and 15% testing data. For the CNN neural, the public application "Application-of-CNN-for-NDVR" was utilized. The average accuracy scores were: training set (93.36%), validation set (90.19%), and testing set (76.56%). For the GAN architecture, we used the system "Keyframes-GAN (IEEE TMM 2023)", resulting in scores of training set (73.16%), validation set (70.05%), and testing set (76.04%). For the proposed approach based on IndRNN, the same application as for the CNN was used but with modifications to the source code architecture. The resultant average key frame identification accuracy scores were: training set (94%), validation set (89.04%), and testing set (86.18%) (in 1.13 times better result than for CNN and GAN). However, the results of this architectural innovation are evident. Evaluated on the UCF-101 dataset, a benchmark for video processing tasks, our system demonstrated impressive performance. This performance not only underscores the effectiveness of the modified IndRNN connections and the attention mechanism but also emphasizes the potential of integrating mathematical tools, such as standard deviation, to enhance the system's accuracy. The combination of modified IndRNN, attention, and our unique enhancement approach offers a fresh perspective on the quest for efficient key frame detection.

V. FUNDING

The National Research Foundation of Ukraine funded this research under project number 2021.01/0103.

VI. CONCLUSIONS

The analysis commences with a comparative study of data processing techniques applied to video content. It is determined that the optimal strategies are grounded in artificial intelligence and machine learning algorithms. Specifically, methods modeling temporal dependencies across a variable range using convolutional neural networks (CNN) are distinguished. The integration of these methods with special "attention" mechanisms has shown a significant prospect in improving accuracy and efficiency.

Furthermore, for the processing of "uncontrolled" video sequences, generative adversarial networks (GANs) coupled with "attention" and "actor-critic" mechanisms proves to be beneficial.

A significant highlight of the study is the proposed approach which employs modified Independent Recurrent Neural Networks (IndRNN). This is integrated with an attention mechanism and enhanced using mathematical tools, notably the standard deviation, to optimize key frame detection. This innovative strategy not only underscores the effectiveness of the modified IndRNN connections but also exemplifies the

potential of integrating mathematical instruments in refining the system's accuracy.

Conclusively, the study emphasizes that to simplify the data adaptation procedure across various domains and their associated application scenarios, there is an imperative need to refine generalization models, with a primary focus on enhancing unsupervised training approaches.

References

[1] H. Tang, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion," *Neurocomputing*, vol. 331, pp. 424–433, 2019. https://doi.org/10.1016/j.neucom.2018.11.038.

[2] R. Vázquez-Martín, and A. Bandera, "Spatio-temporal feature-based keyframe detection from video shots using spectral clustering," *Pattern Recognition Letters*, vol. 34, issue 7, pp. 770–779, 2013. https://doi.org/10.1016/j.patrec.2012.12.009.

[3] Z. Qu, et al., "An improved keyframe extraction method based on HSV color space," *Journal of Software*, vol. 8, issue 7, pp. 1751–1758, 2013. https://doi.org/10.4304/jsw.8.7.1751-1758.

[4] C. Lv, "Key frame extraction for sports training based on improved deep learning," *Scientific Programming*, ed.Muhammad Usman, vol. 2021, 2021, pp. 1–8. https://doi.org/10.1155/2021/1016574.

[5] Y. Yuan, et al., "Spatiotemporal modeling for video summarization using convolutional recurrent neural network," *IEEE Access*, vol. 7, pp. 64676–64685, 2019. https://doi.org/10.1109/ACCESS.2019.2916989.

[6] E. Apostolidis, et al., "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, issue 8, pp. 3278–3292, 2021. https://doi.org/10.1109/TCSVT.2020.3037883.

[7] A. Graves, and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, issue 5–6, pp. 602–610, 2005. https://doi.org/10.1016/j.neunet.2005.06.042.

[8] K. Zhou, et al., "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, issue 1, pp. 7582-7589, 2018. https://doi.org/10.1609/aaai.v32i1.12255.

[9] J. Law-To, et al., "Robust voting algorithm based on labels of behavior for video copy detection," *Proceedings of the 14th ACM International Conference on Multimedia*, 2006, pp. 835–844. https://doi.org/10.1145/1180639.1180826.

[10] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo and B. Guo, "Unsupervised extraction of video highlights via robust recurrent auto-encoders," *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 4633-4641. https://doi.org/10.1109/ICCV.2015.526.

[11] B. Mahasseni, et al., "Unsupervised video summarization with adversarial LSTM networks," *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2982–2991. https://doi.org/10.1109/CVPR.2017.318.

[12] E. Apostolidis, et al., "Video summarization using deep neural networks: A survey," *Proceedings of the IEEE*, vol. 109, issue 11, pp. 1838–1863, 2021. https://doi.org/10.1109/JPROC.2021.3117472.

[13] S. M. Tirupathamma, "Key frame based video summarization using frame difference," *International Journal of Innovative Computer Science & Engineering*, vol. 4, no. 3, pp. 160-165, 2017.

[14] S. Jadon and M. Jasim, "Video Summarization using Keyframe Extraction and Video Skimming," EasyChair Preprint, no. 1181, version 2, pp. 1-5, 2020. https://doi.org/10.1109/ICCCA49541.2020.9250764.

[15] S. Lal, et al. «Online video summarization: Predicting future to better summarize present," *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 471–480. https://doi.org/10.1109/WACV.2019.00056.

[16] M. Elfeki and A. Borji, "Video summarization via actionness ranking," *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2019, pp. 754-763, https://doi.org/10.1109/WACV.2019.00085.

[17] K. Cho, et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2014, pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179.

[18] Mahasseni, Behrooz, et al., "Unsupervised video summarization with adversarial LSTM networks," *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 202–211. https://doi.org/10.1109/CVPR.2017.318.

[19] E. Apostolidis, et al., "A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization," *Proceedings of the 1st ACM International Workshop on AI for Smart TV Content Production, Access and Delivery*, 2019, pp. 17–25. https://doi.org/10.1145/3347449.3357482.

[20] X. He, et al., "Unsupervised video summarization with attentive conditional generative adversarial networks," *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2296–2304. https://doi.org/10.1145/3343031.3351056.

[21] S. Li, et al., "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," 13 March 2018, https://doi.org/10.48550/ARXIV.1803.04831.

[22] H.-T. Nguyen, and T.-O. Nguyen, "Attention-based network for effective action recognition from multi-view video," *Procedia Computer Science*, vol. 192, pp. 971–980, 2021. https://doi.org/10.1016/j.procs.2021.08.100.

[23] Y. Zhang, et al., "Unsupervised object-level video summarization with online motion auto-encoder," *Pattern Recognition Letters*, vol. 130, pp. 376–385, 2020. https://doi.org/10.1016/j.patrec.2018.07.030.

[24] A. Nasreen, K. Roy, K. Roy, G. Shobha, "Key frame extraction and foreground modelling using K-means clustering," *Proceedings of the International Conference on Computational Intelligence, Communication Systems and Networks (CICSYN)*, Latvia, 2015, pp. 141-145. https://doi.org/10.1109/CICSyN.2015.34.

[25] M. Gygli, H. Grabner, L. Van Gool, "Video summarization by learning submodular mixtures of objectives," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3090–3098. https://doi.org/10.1109/CVPR.2015.7298928.

[26] M. Liu, H. Liu, C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017. https://doi.org/10.1016/j.patcog.2017.02.030.

[27] Z. Wang, Y-J. Cha, "Unsupervised deep learning approach using a deep auto-encoder with a one-class support vector machine to detect damage," Structural Health Monitoring, vol. 20, issue 1, pp. 406-425, 2021. https://doi.org/10.1177/1475921720934051.

***NATALYA SHAKHOVSKA,** Doctor of science, Lviv Polytechnic National University, Ukraine, the head of artificial intelligence department, Lviv Polytechnic National University, Ukraine. Research interests: Big data mining, data-warehouses, intelligent systems.*

***NATALIA MELNYKOVA,** Doctor of sciences at Lviv Polytechnic National University, Associate Professor of the Artificial Intelligent Systems Department, Lviv Polytechnic National University, Ukraine, Data Mining, Support Decision Making Systems, Methods of Analysis and Processing of Medical Data.*

***PETRO POBEREIKO,** a Postgraduate student of the Department of Artificial Intelligence Systems of Lviv Polytechnic National University, Lviv Polytechnic National University, Ukraine, Computer vision, Research of Methods of Analyzing Video Flow, Data Mining, Unsupervised Machine Learning.*

***MARYANA ZAKHARCHUK,** an Associate Professor at the department of applied linguistics, vice-director at the Institute of Administration, State Management and Professional Development Lviv Polytechnic National University, Data Mining, Support Decision Making Systems.*