Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Two-Stage Frame Extraction in Video Analysis for Accurate Prediction of Object Tracking by Improved Deep Learning

R. Rajesh Sharma

Computer Science and Engineering, School of Electrical Engineering and Computing, Adama Science and Technology University, Adama, Nazret, Ethiopia

E-mail: sharmaphd10@gmail.com

Abstract

Recently, the information extraction from graphics and video summarizing using keyframes have benefited from a recent look at the visual content-based method. Analysis of keyframes in a movie may be done by extracting visual elements from the video clips. In order to accurately anticipate the path of an item in real-time, the visible components are utilized. The frame variations with low-level properties such as color and structure are the basis of the rapid and reliable approach. This research work contains 3 phases: preprocessing, two-stage extraction, and video prediction module. Besides, this framework on object track estimation uses the probabilistic deterministic process to arrive at an estimate of the object. Keyframes for the whole video sequence are extracted using a proposed two-stage feature extraction approach by CNN feature extraction. An alternate sequence is first constructed by comparing the color characteristics of neighboring frames in the original series to those of the generated one. When an alternate arrangement is compared to the final keyframe sequence, it is found that there are substantial structural changes between consecutive frames. Three keyframe extraction techniques based on on-time behavior have been employed in this study. A keyframe extraction optimization phase termed as "Adam" optimizer, dependent on the number of final keyframes is then introduced. The proposed technique outperforms the prior methods in computational cost and resilience across a wide range of video formats, video resolutions, and other parameters. Finally, this research compares SSIM, MAE, and RMSE performance metrics with the traditional approach.

Keywords: Deep learning, object tracking, Video object prediction, key frame prediction, optimization

DOI: https://doi.org/10.36548/jiip.2021.4.004
Received: 30.10.2021, received in revised form: 28.11.2021, accepted: 11.12.2021, published: 22.12.2021

1. Introduction

Digital cameras, particularly those found in smartphones, and better network topologies have all contributed to the proliferation of video in recent years. Recent figures show that a significant amount of video are being posted on the Internet every minute, and this trend is projected to continue for the foreseeable future, as demand for video content continues to climb [1-5]. The video indexing, archiving, and retrieval systems face major issues as a result of this growth.

The semantic content of these movies must be automatically classified in order to utilize and manage them properly. As a result, the process of categorizing video material remains a difficult one. People's need for video materials has also risen as information technology has advanced through time. Because of this, data and information processing has become a major research topic in relevant units.

Machine learning algorithms are becoming more significant in the present stage of data analysis, and deep learning is one of the most popular information processing technologies [6]. Deep learning approaches are novel technique derived from conventional neural networks. A neural network model is also used for prediction and classification and may be used in a wide range of analytical settings. Deep neural networks outperform traditional neural networks when it comes to classification and model training, particularly when it comes to enhancing server performance and processing data. This technology has been employed in the realms of music and video as well as text and images [7].

Inspired by biological neurons, neural networks may trigger neurons in the brain via machine learning. Relevant and critical information extraction is a must in sports video analysis for action detection. It is difficult to analyze a huge video in a short period of time while maintaining its semantics [8-13]. As a result, one of the first things to do when analyzing video is to extract keyframes. An elegant summary of the full movie sequence may be found in the keyframe. For a specific computer vision job, a video is generally produced at 30 frames per second and provides extra information. Summarization and video localization are the two most common uses of keyframe detection. Video frames will demand additional processing power and memory.

Thus, specialists and academics have started to pay more attention to the application value of deep neural networks, such as the progressive application of deep neural networks to video, audio and text [14]. Recurrent neural networks, for example, have been extensively

employed in in-text processing because of their ability to handle recursive temporal information. Furthermore, the use of convolutional neural networks to classify individual video frames has greatly improved video categorization. More robust video processing performance will be achieved by combining recurrent neural networks with convolutional neural networks, which can now process many frames at once [15].

Researchers in the past have developed video classification algorithms that are based on image classification algorithms, treating each frame of a video as an image and extracting its characteristics in order to get an array of real numbers using a bag-of-words model of fixed length. As part of the training process, Support Vector Machines (SVMs) are used to identify additional video frames.

2. Research Structure

This research structure has been arranged as follows; section 3 provides the history of the past research work about video frame tracking; section 4 delivers the proposed research work, the overcome of past research work. The experimental test has been conducted and discussed in section 5. Finally, it concludes with future possible research work.

3. Related Works

Content-based video retrieval has grown significantly as a result of advances in science and technology. Video-related material given by the user as the search criterion is used to search the video library for comparable videos. For face clustering in news videos, Anantharajah et al. used a similar metric-based quality evaluation approach. With the face quality metrics-based technique, high-speed real-time processing may be achieved [16].

An in-depth CNN-based face quality evaluation for video selection was suggested by Vignesh et al. in their proposal. They were able to surpass the rank-based technique without using any pre-defined face characteristics since they made use of DNN's outstanding feature learning capacity [17].

An integrated system solution based on video content information parsed was offered by Zhong and Smoliar. After analyzing a video sequence, a group of frames that best reflect the scene's visual information is retrieved by humans in contemporary video indexing systems [18].

In the video indexing process, these frames are known as crucial frames. Using low-level data like color histograms and motion information, Zhang et al. have developed a feasible way for keyframe extraction [19].

Changing the threshold for recognizing "significant" color histogram changes and the overlap ratio of key-frames in panning sequences, allows the user to alter the density of key-frames or the abstraction ratio. The video's real content will dictate the number of crucial frames that can be extracted from it in the final product. Hanjalic et al. claim that this is a drawback of this kind of key-frame extraction [20].

The Keyframe extraction has been used to create a video overview of the moving mechanism. Each frame's visual attention is measured using a description known as the attention quantifier, which signals color conspicuity and motion with higher attention. Using L1-norm and cumulative optical fluxes, Abdullah et al. suggested an action keyframe extraction approach. For salient region-based keyframe extraction, researchers have used optical flow and calculated mutual information entropy in a similar manner. Analyzing manually annotated data is the foundation of most existing frame annotation techniques. Keyframes annotated by a diverse group of people are often used in video summarizing methods, such as the most popular critical frame-based techniques. Video annotation and crucial frame identification, on the other hand, account for the majority of the related effort [21].

3.1 Research gap

Keyframe selection on video streams is a potential approach for overcoming time and resource constraints. Previously, the content of a scene may be represented by a keyframe in partial. This research article has developed a two-stage key frame extraction with Convolutional Neural Network (CNN) keyframe extraction engine that extracts keyframes in accordance with face quality, lowering the data volume and supplying keyframes with high-quality faces for face recognition.

4. Proposed Method

4.1 Video Frame Extraction

The extraction of keyframes from any digital video format is an essential aspect of this study. To enable viewers to experience the video material in a unique way, the grey value is

set to a certain range. Four low-gray pictures have been taken from a single movie, making it impossible for most viewers to discern the content. Figure 1- 3 contains the details of this entire research framework.

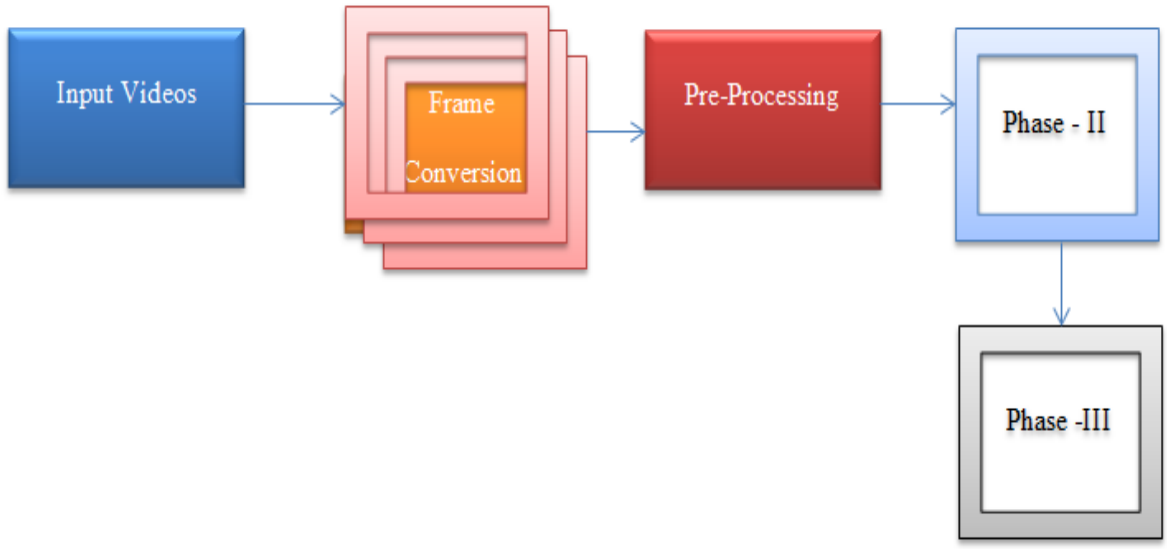

Figure 1. Pre-processing steps of proposed framework

4.2 Proposed Modified Object Prediction Technique

Learning by prediction is used to test the accuracy of the system's knowledge of the underlying patterns in the input data under the premise that good predictions can only arise from accurate representations. So it's a useful framework for learning representations. Predicting future outcomes based on previous inputs is the heart of the predictive learning paradigm. Thus, video prediction involves forecasting future frames in an ongoing video, given the context of the past frames in the movie. This is how video prediction works. Video prediction differs from video generation such that, it is based on a previously learnt representation of a series of input frames, as opposed to video creation, which is mostly unconditioned [22-25].

This challenge is considered as a supervised learning strategy since the target image is used as the label for each predicted future video frame. No additional tagging or human monitoring is required, as this information is already included in the video sequence. In this way, prediction-based learning fills the gap between supervised and unsupervised methods of instruction.

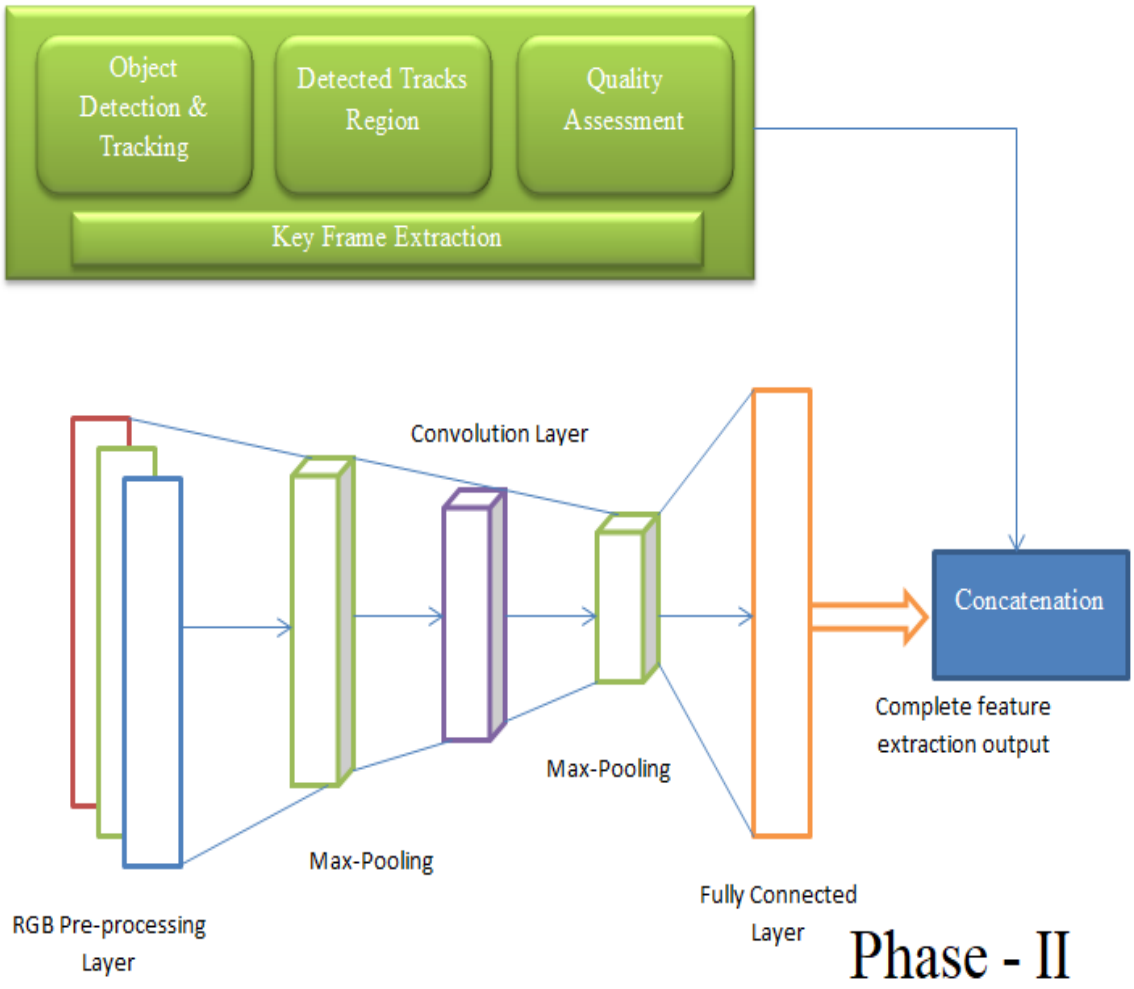

Figure 2. Two-stage Feature Extraction Algorithm

4.2.1 CNN approach for video frame

The spatial structure of pictures is effectively modelled by convolutional neural networks (CNNs), which are the fundamental building blocks of deep learning architectures optimized for visual reasoning [28]. The predictive learning literature that focuses on visual prediction has its roots in CNNs. However, the inter and intra-frame dependencies restrict their performance. Due to the restricted receptive fields of the kernel size, short-range intraframe dependencies are accounted via convolutional techniques [26-29]. The stages listed below are included in this study:

Step 1

Increasing the number of convolutional layers, increasing the kernel size, or mixing different scales linearly, as in the reconstruction of a Laplacian pyramid, are the possible ways to increase the complexity of the structure.

Step 2

Dilated convolutions are used to capture long-range spatial relationships, while the max-pooling operation is used to lower the size of an array by the same amount.

Step 3

Increasing the receptive fields using subsampling: In exchange for sacrificing resolution, this is accomplished by the use of a pooling process. The modification of the slight changes in CNN addresses the problem of sparse view representation by using residual connections to maintain resolution while increasing the number of stacking convolutions used.

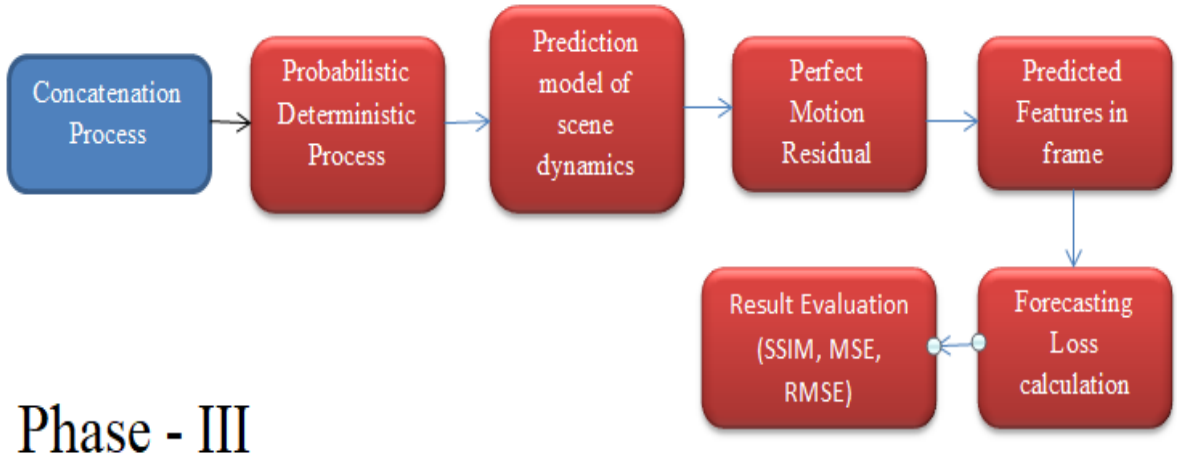

Figure 3. Forecasting motion prediction for object tracking

4.2.2 Improving proposed techniques

In order to anticipate future precipitation intensity in a particular region, precipitation nowcasting is seen as a video prediction issue using a fixed camera and weather radar. For precipitation forecasting, ConvLSTM outperforms when fully linked with the proposed neural network in terms of spatial domain approach through neighbor pixel correlation.

An essential component of intelligence and one of the primary aims of decision-making systems is the ability to foresee and reason about future occurrences. There are biological foundations for this concept, and it takes influence from the cognitive neuroscience field's predictive coding paradigm. According to neuroscience, the human brain is capable of creating intricate conceptual models of the physical and causal principles that govern our universe. Observation and interaction are the primary methods [30, 31] to achieve this. As we get older, we learn concepts like biological motion and intuitive physics, among other things, which help

us make sense of the world around us. The brain continually learns and refines the world models that it already understands via prediction. Early attempts such as [22], [28], and [30] attempted to build computer models that captured the essence of the predictive coding paradigm. The prediction of the frames in video is regarded as significant since it accurately reflects the principles of predictive coding.

4.2.3 Evaluation of image quality

The Structural Similarity Index Measurement (SSIM) for image quality evaluation frames are compared and the similarity of the two pictures is measured using a value closer to 1, indicating a more comparable quality between the images. According to the notion of structural similarity, there is a substantial association between pixels, particularly in the airspace nearest to each other, which carries crucial information about the organization of the picture. Structural information may be extracted from the view using the human visual system, which can be used to gauge perceived picture quality [12].

4.2.4 Importance of Structural Similarity Measure

According to the color feature key, this issue is unaffected by eliminating the extraction of frame structure information. The program's structure is based only on an index of similarity, much like the components themselves. The notion of the structural similarity index defines structural information as a component of an image composition apart from brightness and contrast. Object characteristics might be reflected in it as well. To measure structural similarity, the covariance is used in this research work.

4.3 Optimization for the Proposed Framework

The video prediction algorithms that were developed earlier sought to forecast future pixels of the image or frames in the video analysis, which were more focused on the full object moving part. Using k-means, it clustered video frames into patch groups. In a k-means discretized space, it was believed that all non-overlapping patches were equally distinct, yet commonalities might be detected between them. Short-term predictions are made using a convolutional extension of an RNN-based model. The full-resolution frame shows some impact of tilling since it is composed of the expected patches. Large and fast-moving objects are accurately predicted. However, tiny and slow-moving items still have an opportunity for development.

5. Results and Discussion

The experimental result has been discussed in this section. The annotated YouTube videos make up the Sports1M dataset, which is a video categorization dataset. There are about 400 classes in this instance, all of which have the sports label extracted from the YouTube Topics API. Table 1 contains the performance metrics computation.

Table 1. Computation of performance metrics

Video Prediction Techniques	Video Type	Total frame	Extracted key frames	Processing time of Each frame	MAE	RMSE	SSIM
Key Frame Extraction Method	mp4	789	109	0.019s	19.45	30.82	0.73
Forecasting Prediction method			77	0.142s	21.07	31.76	0.53
Proposed Two-stage frame extraction			178	0.048s	14.52	28.24	0.81

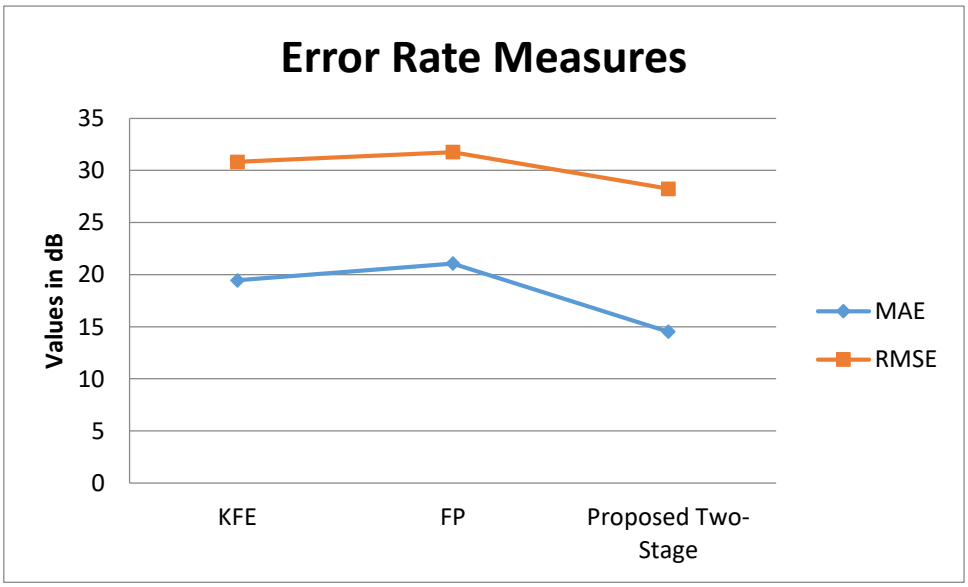

Figure 4. Entire error rate measurement comparison

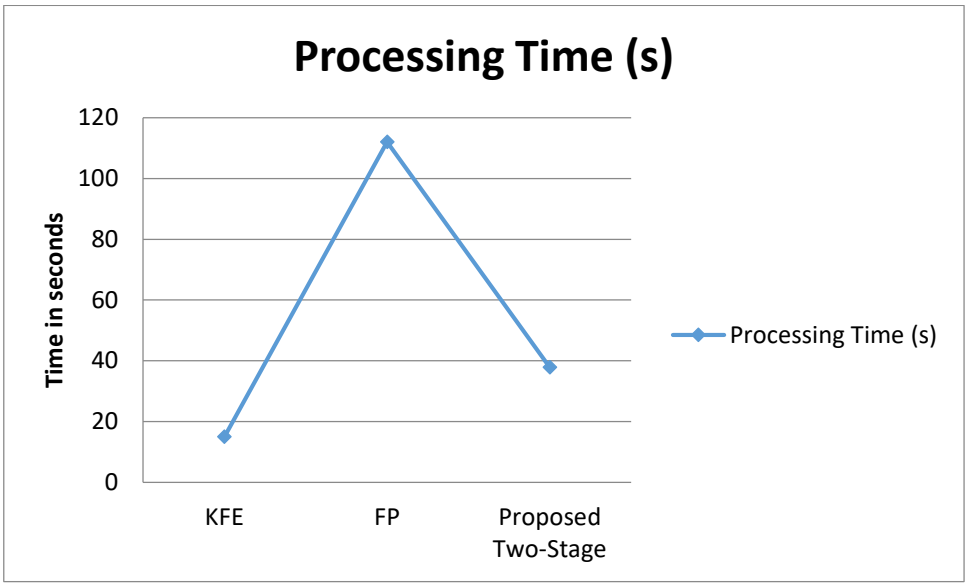

Figure 5. Processing time comparison

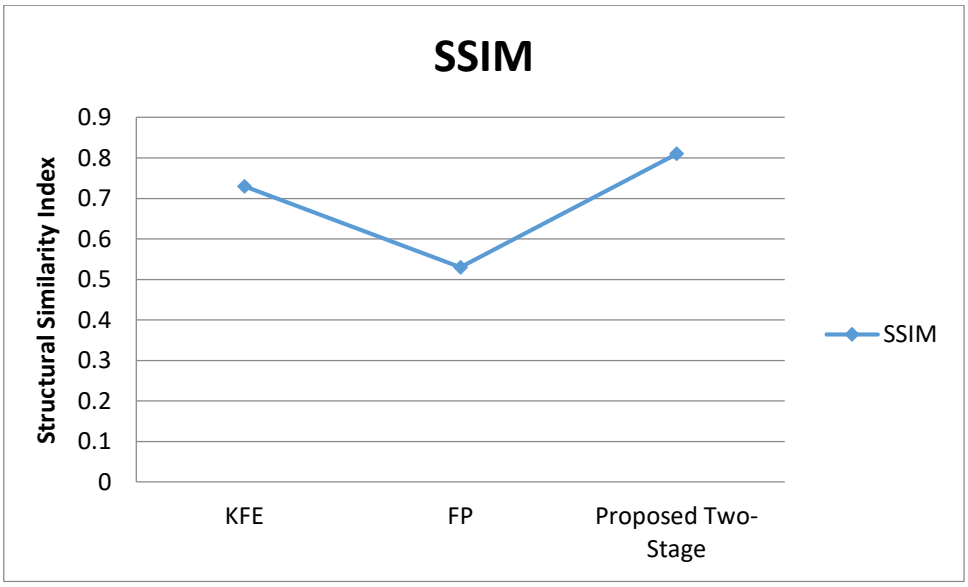

Figure 6. Structural similarity index measurement comparison

When a video is accessed through YouTube, the video's quality, length, and frame rate can be adjusted. The bouncing ball dataset is a typical test set for high-dimensional sequence models. It's a simulation of three balls bouncing in a container. The usual structure includes 4000 training videos, 200 testing films, and 200 more for validation.

Figure 4-6 depicts the total process performance analysis using the standard benchmark dataset and metrics as a starting point for comparison. Its clips are created randomly with variable resolution. When compared to the frame prediction (FP) approach, the Key Frame Extraction (KFE) method exhibits a modest error rate. However, as compared to the other two

techniques, the suggested two-stage procedure includes less Mean Absolute Error (MAE) values. Furthermore, the suggested approach outperforms the two classic algorithms in terms of additional performance metrics such as processing time and SSIM. These datasets are solely devoted to the task of making predictions about videos. This investigation used a sample 1.mp4 movie of 789 frames which is a brief video of sports and bouncing ball prediction.

Mean Absolute Error (MAE) and RMSE metrics performs well for the suggested two-stage frame extraction approach, as measured by SSIM for the projected frames. However, the processing time is a little concerning, especially when considering the lengthy duration of the complete video processing procedure.

6. Conclusion

This study report shows that current approaches in video retrieval have been thoroughly studied to offer related enhancement methods, such as Key frame and feature extraction. In addition, a vital frame extraction approach for video frame tracking protection has been provided and obtained using low-level characteristics and frame difference comparisons. It is planned to use more complex neural network models and refine the model's structure in the future to increase face quality assessment performance in accuracy and speed. On two separate datasets, the suggested crucial frame extraction approach has outperformed the generally utilized keyframe extraction methods. Future research might concentrate on adapting the suggested approach to real-world video classification and video summarization challenges utilizing more sophisticated architectures.

References

[1] Raj, Jennifer S., and Mr C. Vijesh Joe. "Wi-Fi Network Profiling and QoS Assessment for Real Time Video Streaming." IRO Journal on Sustainable Wireless Systems 3, no. 1 (2021): 21-30.

[2] C. Huang and H. Wang, “Novel key-frames selection framework for comprehensive video summarization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, pp. 577–589, 2019.

[3] Sungheetha, Akey, and Rajesh Sharma. "Real Time Monitoring and Fire Detection using Internet of Things and Cloud based Drones." Journal of Soft Computing Paradigm (JSCP) 2, no. 03 (2020): 168-174.

[4] M. Jian, S. Zhang, L. Wu, S. Zhang, and X. Wang, “Deep key frame extraction for sport training,” Neurocomputing, vol. 328, pp. 147–156, 2019.

[5] Sharma, Rajesh, and Akey Sungheetha. "An Efficient Dimension Reduction based Fusion of CNN and SVM Model for Detection of Abnormal Incident in Video Surveillance." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 55-69.

[6] S. Wen, W. Liu, Y. Yang, T. Huang, and Z. Zeng, “Generating realistic videos from keyframes with concatenatedGANs,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, pp. 2337–2348, 2019.

[7] Chen, Joy Iong Zong, and P. Hengjinda. "Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method-A Comparative Study." Journal of Artificial Intelligence 3, no. 01 (2021): 17-33.

[8] L. Wu, J. Zhang, and F. Yan, “A pose let based key frame searching approach in sports training videos,” in Proceedings of the Information Processing Association Annual Summit and Conference, pp. 1–4, Hollywood, CA, USA, December 2012.

[9] Raj, Jennifer S., and J. Vijitha Ananthi. "Recurrent neural networks and nonlinear prediction in support vector machines." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 33-40.

[10] M. Jian, S. Zhang, L. Wu, S. Zhang, X. Wang, and Y. He, “Deep key frame extraction for sport training,” Neurocomputing, vol. 328, pp. 607–616, 2018.

[11] Mugunthan, S. R., and T. Vijayakumar. "Design of Improved Version of Sigmoidal Function with Biases for Classification Task in ELM Domain." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 70-82.

[12] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” TPAMI, vol. 40, no. 4, 2018.

[13] Vijayakumar, T., Mr R. Vinothkanna, and M. Duraipandian. "Fusion based Feature Extraction Analysis of ECG Signal Interpretation–A Systematic Approach." Journal of Artificial Intelligence 3, no. 01 (2021): 1-16.

[14] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, “Understanding the Effective Receptive Field in Deep Convolutional Neural Networks,” in NeurIPS, 2016.

[15] Sathesh, A., and Edriss Eisa Babikir Adam. "Hybrid Parallel Image Processing Algorithm for Binary Images with Image Thinning Technique." Journal of Artificial Intelligence 3, no. 03 (2021): 243-258.

[16] K. Anantharajah, S. Denman, D. Tjondronegoro, S. Sridharan, C. Fookes, and X. Guo. Quality based frame selection for face clustering in news video. In Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on, pages 1–8. IEEE, 2013.

[17] S. Vignesh, K. M. Priya, and S. S. Channappayya. Face image quality assessment for face selection in surveillance video using convolutional neural networks. In Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on, pages 577–581. IEEE, 2015.

[18] Zhang HJ, Wu J, Zhong D, Smoliar SW (1997) An integrated system for content-based video retrieval and browsing. Pattern Recognit 30(4):643–658.

[19] H. J. Zhang, C. Y. Low, and S.W. Smoliar, "Video parsing and browsing using compressed data", Multimedia Tools Appl. 1, 1995, 91–113.

[20] A. Hanjalic and R.L. Langendijk, "A New Key-Frame Allocation Mathod for Representing Stored Video Streams", Proc. of 1st Int. Workshop on Image Databases and Multimedia Search, 1996.

[21] Abdullah SNHS, Ng KW (2017) Action key frames extraction using L1-norm and accumulative optical flow for compact video shot summarisation. In: Advances in visual informatics: 5th international visual informatics conference, IVIC 2017, Bangi, Malaysia, November 28–30, 2017, proceedings, vol 10645. Springer, p 364.

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016.

[23] Karuppusamy, P. "Building Detection using Two-Layered Novel Convolutional Neural Networks." Journal of Soft Computing Paradigm (JSCP) 3, no. 01 (2021): 29-37.

[24] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, “Decomposing motion and content for natural video sequence prediction,” in ICLR, 2017.

[25] Pandian, A. Pasumpon. "Performance Evaluation and Comparison using Deep Learning Techniques in Sentiment Analysis." Journal of Soft Computing Paradigm (JSCP) 3, no. 02 (2021): 123-134.

[26] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, “Eidetic 3d LSTM: A model for video prediction and beyond,” in ICLR, 2019.

[27] Jain, Sarika, Ekansh Tiwari, and Prasanjit Sardar. "Soccer Result Prediction Using Deep Learning and Neural Networks." In Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020, pp. 697-707. Springer Singapore, 2021.

[28] Dhamodaran, S., Ch Krishna Chaitanya Varma, and Chittepu Dwarakanath Reddy. "Weather Prediction Model Using Random Forest Algorithm and GIS Data Model." In International Conference on Innovative Data Communication Technologies and Application, pp. 306-311. Springer, Cham, 2019.

[29] Asha, J., S. Rishidas, S. SanthoshKumar, and P. Reena. "Analysis of Temperature Prediction Using Random Forest and Facebook Prophet Algorithms." In International Conference on Innovative Data Communication Technologies and Application, pp. 432-439. Springer, Cham, 2019.

[30] Rani, Pooja, Rajneesh Kumar, and Anurag Jain. "Multistage Model for Accurate Prediction of Missing Values Using Imputation Methods in Heart Disease Dataset." In Innovative Data Communication Technologies and Application, pp. 637-653. Springer, Singapore, 2021.

[31] Ishi, Manoj S., and J. B. Patil. "A Study on Machine Learning Methods Used for Team Formation and Winner Prediction in Cricket." In Inventive Computation and Information Technologies, pp. 143-156. Springer, Singapore, 2021.

Author's biography

R. Rajesh Sharma is a Computer Vision and Robotics SIG Coordinator in the department of Computer Science and Engineering in School of Electrical Engineering and Computing in Adama Science and Technology University, Adama, Nazret, Ethiopia. He has more than 8 years of academic experience. His areas of research are networking, probabilistic computing, fuzzy, bio- inspired computing, data visualization, fault diagnosis, robotics, internet of things, neurocomputing, information retrieval, human-machine interface and network security. He has published more than 20 international and national journals. He is a life member of International Association of Engineers and Indian Society of Technical Education, and member of IAENG, IACSIT and AACE.