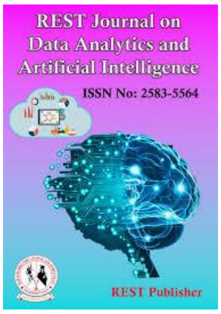
REST Journal on Data Analytics and Artificial Intelligence

Vol: 3(3), September 2024

REST Publisher; ISSN: 2583-5564

Website: https://restpublisher.com/journals/jdaai/

DOI: https://doi.org/10.46632/jdaai/3/3/6



A Comprehensive Study on Advanced Video Data Preprocessing Techniques for Enhanced Object Detection

***Roopa R, Humera Khanam**

Sri Venkateshwara University College of Engineering, Tirupati, Andhra Pradesh, India.

*Corresponding Author Email: roopa509@gmail.com

Abstract: Video processing has become a vital area in computer vision and deep learning, with diverse applications including crowd analysis, anomaly identification, and activity tracking. Although numerous surveys have examined various aspects of these functionalities, there is still a requirement for a complete review that combines these findings into a coherent perspective. This survey study provides a comprehensive analysis of several model architectures, emphasising their advantages, shortcomings, and constraints. We also emphasise the profound influence of these technologies in several fields, such as surveillance, healthcare, and autonomous systems, specifically focussing on the applications of deep learning in video processing. Our review not only analyses the latest advancements but also explores the complex processes and tactics used by deep learning models to derive valuable insights from video data. Furthermore, we examine the importance of accessible datasets and their crucial role in propelling research progress in this field. By outlining the obstacles and concerns that researchers have while adopting these systems, we offer a clear plan for future research paths. We want to stimulate ongoing innovation and advancement in the domain of video processing using deep learning techniques.

Keywords: Video Processing, Deep Learning, object detection, Model Architectures, Surveillance Technology.

1. INTRODUCTION

Deep Learning (DL) techniques have been extensively used in diverse fields such as speech recognition, image classification, and text detection. Nevertheless, due to the rapid increase in video content, social media platforms like Facebook, Twitter, Instagram, and TikTok are placing more emphasis on enhancing their services with video-oriented features. The ongoing uploading and dissemination of countless videos on these platforms have generated substantial apprehensions surrounding privacy, suitability of material, and the widespread presence of counterfeit, aggressive, and unsuitable videos. This has emphasised the pressing requirement for efficient content filtering and takedown procedures.

In order to tackle these difficulties, it has become crucial to develop real-time video processing and recognition systems to improve automated security measures. Scientists are currently engaged in efforts to enhance AI-powered video processing in order to fulfil the increasing need for dependable and effective content moderation. Nevertheless, even with the high demand for it, the process of automatically identifying video content remains intricate and necessitates accurate methods and creative solutions. Comprehending the core concept of video is essential in this context. A video is a series of digital images that are processed quickly, with each image referred to as a frame per second. Presently, the predominant area of study is centred around particular situations in which sequences of images are examined as a component of computer vision objectives.

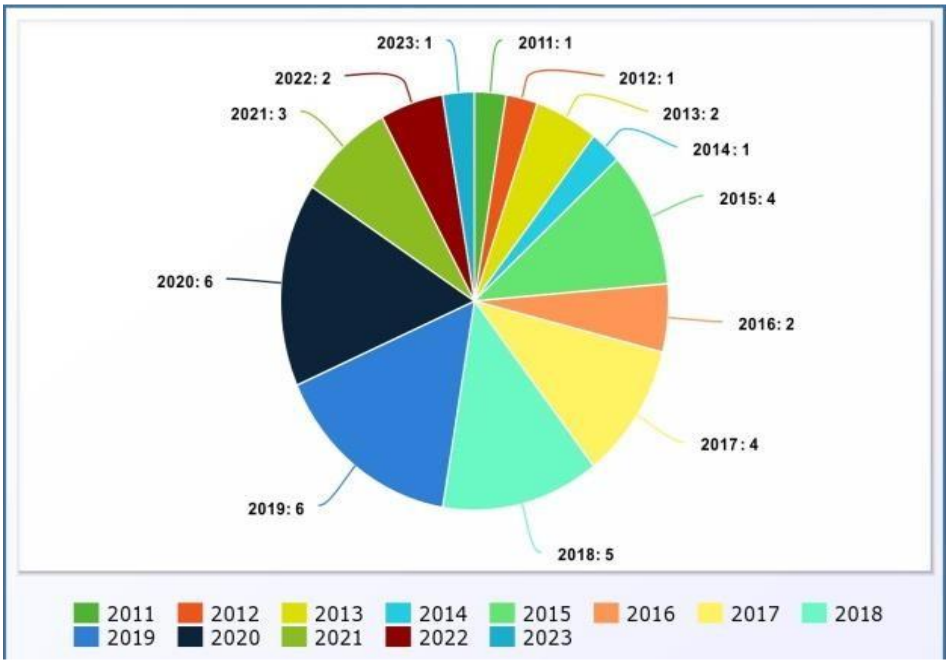


FIGURE. 1. Year wise paper used in this survey

Figure 1 displays a pie chart that depicts the distribution of research papers utilised in this study, categorised based on their year of publication. This visual depiction offers a concise summary of the distribution of research contributions throughout various years, emphasising the temporal patterns in the exploration of video data preprocessing and deep learning applications. The segments of the pie chart reflect different years, and the size of each segment corresponds to the proportion of articles published in that year compared to the total number of papers assessed in this survey. This visualisation facilitates the rapid identification of years with substantial research output, suggesting periods of important developments or heightened interest in video processing techniques.

By analysing the optical flow in a sequence of photographs, it is feasible to trace the path of an item and anticipate its future movements. Expanding on this comprehension, our research investigates the notion of video processing with the objective of recognising and explaining temporal and spatial characteristics within visual data. The utilisation of many modes of input is very beneficial for tasks like as identifying objects, tracking motion, and detecting events. The integration of video material on social media platforms necessitates the advancement of automated, frame-by-frame video analysis to improve content control and security.

2. TECHNIQUES, MODELS &DATASET FOR VIDEO PREPROCESSING

The increasing presence of video material on many platforms has created a demand for sophisticated approaches to preprocess video data. With the rise of video as a prominent means of communication, entertainment, and information dissemination, it is crucial to prioritise the excellence and comprehensibility of video content. Video preprocessing is an essential initial stage in the video processing pipeline, establishing the groundwork for more intricate analyses including object detection, motion tracking, and event recognition. The objective of video preprocessing is to optimise the visual quality of raw video data by minimising noise and preparing it for subsequent processing activities. By utilising these preprocessing approaches, the effectiveness and precision of subsequent processes are greatly enhanced, resulting in superior overall performance of video-based applications.

This section delves into various fundamental strategies that have been devised to enhance the efficiency of video data prior to subsequent processing. These strategies tackle a range of difficulties in video processing, including reducing noise and improving specific areas of interest as described by the user. In addition, we analyse various models that have been utilised in video preprocessing, offering a thorough examination of the most advanced techniques now employed in this domain**.**

Techniques:

We identified several key techniques for video data preprocessing that are essential for enhancing and optimizing video content before it undergoes further processing. These techniques aim to improve user viewing experiences, reduce noise, and enhance the overall quality of video playback. Below, we describe the most notable techniques:

- **Frame Extraction and Edge Detection:** Extracting individual frames from videos allows for detailed processing on a per-frame basis. Edge detection plays a critical role in emphasizing essential features while minimizing noise, thereby improving the clarity of important visual elements.
- **Motion Estimation for Noise Reduction:** By applying motion estimation techniques, noise within the video can be effectively reduced, resulting in clearer and more visually appealing video content.
- **Video Image Decoding:** This technique ensures the quality and consistency of zooming operations within user-defined regions of interest, thereby enhancing the viewer's experience by maintaining image integrity during zoom actions.
- **Smooth Viewpoint Transitions:** In omni-directional videos, smooth transitions between different viewpoints are crucial to avoid abrupt changes in content. This technique ensures a seamless and fluid viewing experience.
- **Automatic Information Placement Detection:** This method automates the detection and optimal placement of information within the video, thereby improving the efficiency and effectiveness of content presentation.
- **Video Clipping Based on Time Periods and Photographic State:** Clipping video data according to specific time periods and photographic states allows for precise selection of relevant segments, which is particularly useful for targeted analysis or specific applications.

Models:

In our research, we identified five major models that have been effectively employed in video preprocessing. Each of these models offers unique advantages and has been applied to various video identification and analysis tasks. Below is an overview of these models, with a summary of their application in research papers presented in Table 1.

Convolutional Neural Networks (CNNs):

CNNs are widely used in video preprocessing due to their ability to accurately detect intricate patterns and features within images. They are particularly effective in tasks such as object identification, where their capability to extract spatial-temporal information from video data is crucial. The author [8] discusses three general connection patterns—early fusion, late fusion, and slow fusion—that enhance the network's temporal connectedness, making it easier to extract meaningful information from video sequences. CNNs have also been employed in event detection within video clips, where a novel method combines spatial and temporal data with a frame descriptor to improve visual content extraction. Additionally, a new approach for Human Activity Recognition (HAR) integrates spatial and temporal networks at the convolutional layer, maintaining high performance without compromise. Another study introduces a multi-DCNN framework for multimodal gesture recognition, combining 3D Convolutional Deep CNN (3D-CDCN) and 2D Multi-Resolution CNN (2D-MRCN).

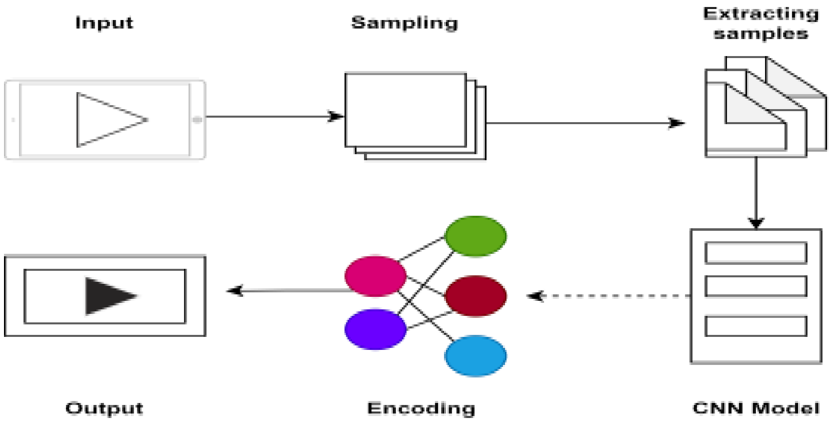


FIGURE 2. An example of video data preprocessing in CNN model

Figure 2 provides a detailed overview of the video data preprocessing workflow within a Convolutional Neural Network (CNN) model. The figure outlines the key stages involved in transforming raw video input into a processed format that can be effectively analyzed by the CNN, ensuring accurate and reliable outputs.

Input: The process begins with the raw video input, which consists of a sequence of frames captured over time. This raw data often contains various challenges such as noise, varying resolutions, and inconsistencies in lighting or motion, all of which need to be addressed through preprocessing.

Sampling: The next step involves sampling the video data, where specific frames or segments are selected based on predefined criteria. This step is crucial for reducing the volume of data while retaining the most relevant portions of the video, ensuring that the subsequent analysis is both efficient and focused.

Extract Samples: Following sampling, the selected frames or segments are extracted. This step isolates the key frames that will be used in the CNN model, allowing for targeted analysis of specific moments within the video. The extracted samples are then prepared for further processing.

CNN Model: The pre-processed frames are fed into the CNN model, where they undergo a series of convolutional operations. These operations involve filtering the frames to detect patterns such as edges, textures, and other critical features. The CNN model's multiple layers progressively refine the data, enabling the network to learn and recognize complex visual structures.

Encoding: After the convolutional layers have processed the frames, the extracted features are encoded into a format suitable for classification or further analysis. This encoding step often involves compressing the feature maps into a lower-dimensional representation while preserving the essential information needed for decision-making.

Output: The final output of the CNN model is a set of predictions or classifications based on the encoded features. This output might include object detection, action recognition, or anomaly detection results, depending on the specific application of the CNN model.

Deep Neural Networks (DNNs):

DNNs are complex artificial neural networks composed of multiple layers of interconnected neurons, allowing them to learn intricate patterns from data autonomously. DNN-based approaches have been applied to various tasks, including road sign and lane detection in autonomous driving [18]. Pashchenko et al. [15] developed a DNN-based model for recognizing critical situations in transport systems, while Amosovetal [26] introduced a technique for classifying probabilities in video fragments, effectively detecting normal and abnormal situations. Another application of DNNs involves video processing for anomaly detection using sparse coding techniques.

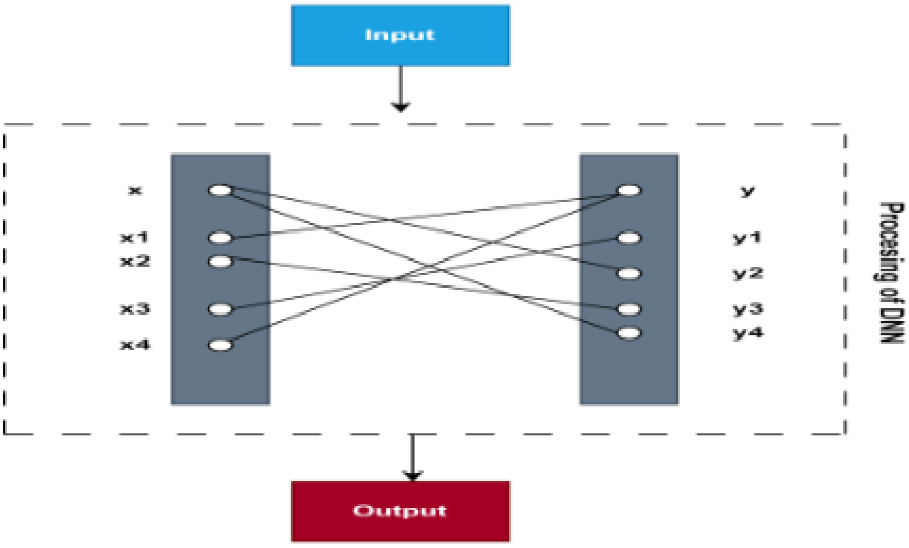


FIGURE 3. Basic DNN architecture

Figure 3 illustrates a core Deep Neural Network (DNN) structure consisting of three main components: the input layer, hidden layers, and output layer. The input layer of the network receives unprocessed data, such as video frames or features, and transmits them into the network. Subsequently, this data traverses via one or more concealed layers, whereby the network's fundamental learning takes place. The hidden layers of the network contain neurones that utilise weights and activation functions to process the input data. This enables the network to acquire intricate patterns and representations. The output layer ultimately generates the network's predictions or classifications, depending on the particular job, such as object detection or video classification. This fundamental structure serves as the basis for numerous deep learning models, allowing the deep neural network (DNN) to effectively process intricate, non-linear connections within data and produce precise forecasts.

Recurrent Neural Networks (RNNs):

RNNs are specialized neural networks designed to analyze sequential input by maintaining a hidden state that stores information from earlier parts of the sequence. Researchers have leveraged RNNs to address challenges such as recognizing faces from low-quality video frames [6]. For object segmentation in video data, a novel technique combines Convolutional RNNs with optical flow, creating masks for earlier frames while isolating items from the background in video sequences. This approach efficiently consolidates information from video frames, leading to significant advancements in video processing. Additionally, RNN-based methods have been developed for detecting video tampering, marking a substantial progression in this field.

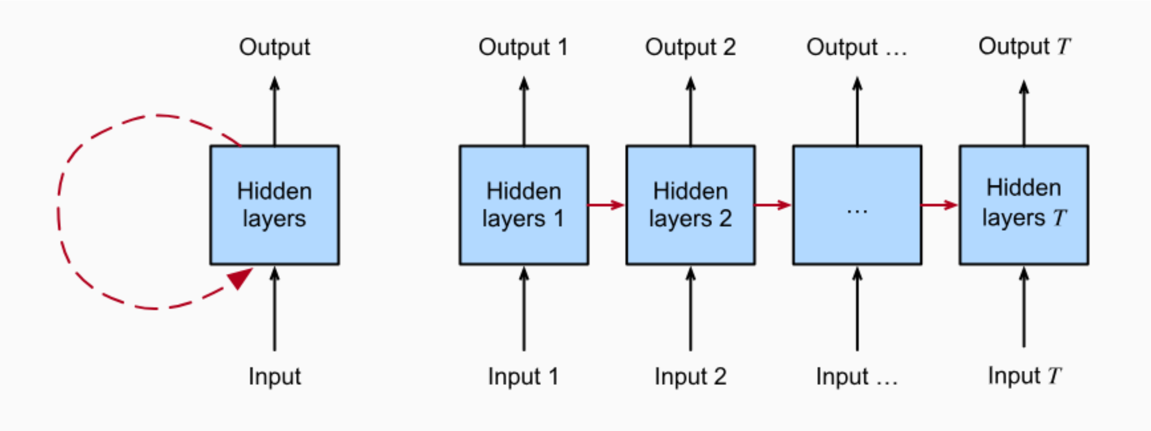


FIGURE 4. RNN-based video data preprocessing architecture [27]

Figure 4 illustrates the construction of a video data preprocessing system based on Recurrent Neural Networks (RNN). The system consists of four main components: the input layer, hidden layers, RNN block, and output layer. The input layer's primary function is to receive consecutive video frames, which act as the unprocessed data for the preprocessing pipeline. Subsequently, the frames are transmitted to the concealed layers, where the recurrent neural network (RNN) block assumes a pivotal function. The RNN block is specifically designed to capture temporal dependencies by preserving a concealed state that retains information from prior frames. This enables the network to efficiently analyse sequential input, comprehending patterns and connections over a period of time. The concealed layers, furnished with the RNN block, extract significant characteristics that depict the temporal dynamics inside the video clip. Ultimately, the processed data is transmitted to the output layer, where the analysis yields outcomes such as classifications or forecasts, which are determined by the acquired temporal patterns.

Hybrid Models:

Hybrid models combine two or more approaches from different domains or perspectives to achieve enhanced performance or tackle complex problems. By integrating the strengths of each component, hybrid models offer more accurate results and improved system capabilities. For example, the authors in [21] proposed using temporal segment networks for human activity recognition. This approach breaks down video data into smaller segments, extracts samples from each segment, classifies them using a multistream CNN network, and then aggregates a softmax score for the entire video. Another study [22] introduced a unique method for recognizing human activity by combining inflated 3D CNNs and LSTM models. The high-level temporal information produced by the pre-trained 3D CNN model on the Kinetics dataset was recorded using Long Short-Term Memory (LSTM) networks.

Transformer-Based Models:

Transformer-based models have gained prominence for their ability to extract critical frames or snippets from videos to create summaries. One such technique involves using a Transformer encoder-decoder network structure [21], where the encoder consists of feed-forward and self-attentional neural network modules. The Transformer model abstracts the video's spatial-temporal structural information, and deep reinforcement learning is applied to enhance the model's learning capacity. These approaches have demonstrated superior performance in generating video summaries compared to existing methods, especially when trained on video abstract datasets.

Table 1 provides a comprehensive summary of various research papers focused on applying different algorithms to specific applications in the field of video processing. The table outlines the algorithm used, the application domain, performance metrics, and the year of publication, offering insights into the effectiveness of these approaches across diverse tasks.

- **CNN-based Approaches:** The Convolutional Neural Network (CNN) is a widely used algorithm across multiple applications such as self-driving car obstacle detection, human activity recognition, gesture detection, and inappropriate content detection. The performance metrics vary, with accuracy rates ranging from 72.53% to 95.66%, and a mean average precision (mAP) of 80% for human activity recognition.
- **DNN-based Methods:** Deep Neural Networks (DNNs) have been applied in transport systems and situation identification, demonstrating high accuracy and precision, particularly in the transport system with a performance of 98.21% accuracy and 93.94% precision.
- **RNN-based Techniques:** Recurrent Neural Networks (RNNs), including variants such as MARN RNN and CRF as RNN, are applied in tasks like face identification and background evaluation. Although these methods show promise, the performance metrics indicate a moderate success rate, with accuracy ranging from 74.32% to 75%.
- **Hybrid Models:** Combining CNNs with LSTMs and other architectures, such as 3D CNNs, has proven effective in human activity recognition and classification, with accuracy rates of up to 83.54%. These hybrid models leverage the strengths of both CNNs and RNNs to improve temporal and spatial feature extraction.
- **Advanced Models and Architectures:** Recent advancements include models like RESNet18 for sign language classification, achieving a remarkable 99.9% accuracy, and ViS4MER for movie clip classification with an 88.4% accuracy. Other novel architectures like DSwarm-Net have shown significant success in human action identification with an accuracy of 98.33%.
- **Transformer-based Approaches:** Transformers, particularly in tasks like transferring vision, have demonstrated strong performance, with accuracy rates approaching 87.8%. These models represent the cutting edge of video processing, providing new avenues for improving real-time video analysis.
- **Miscellaneous Approaches:** The table also includes methods like RANKPOOL+CNN for background evolution, with a precision of 61%, and SVPA for clinical interpreted images, showing an efficiency rate of 98.54%.

TABLE 1. Summary of the papers

Ref.	Algorithm	Application	Performance	Year
[24]	CNN	Self -driving car obstacle detection	Accuracy - 88.6%	2020
[8]	CNN	Reorganization of human activity	mAP-80%	2014
[4]	CNN	Gesture Detection	Accuracy -72.53%	2020
[15]	DNN	Transport system	Accuracy - 98.21% ,Precision -93.94%	2019
[18]	Ensemble DNN	Situation identification	Accuracy - 79.15% ,Precision -69.06%	2019
[6]	MARN RNN	Face identification	Accuracy -74.32%	2019
[17]	CRF as RNN	Background evaluation	-	2018

[3]	CNN+LSTM	Human activity reorganization	Accuracy 83.54%	2017
[21]	3D CNN+LSTM	Human activity identification	Accuracy 83.54%	2017
[5]	RANKPOOL+CNN	Background Evolution	Precision- 61%	2016
[1]	ESN	-	Accuracy 75%	2022
[23]	CNN	Inappropriate content detection	Accuracy 95.66%	2022
[20]	3D CNN	Human activity classification	Precision 87.4%	2022
[12]	SVPA	Clinical interpreted image	Effeciency- 98.54%	2022
[16]	RESNet18	Sign language classification	Accuracy 99.9%	2022
[7]	ViS4MER	Movie clip classification	Accuracy 88.4%	2022
[11]	MVi-Tv2s	-	Accuracy 88.8%	2022
[22]	Transferring vision	-	Accuracy 87.8%	2023
[1]	DSwarm-Net	Human action identification	Accuracy 98.33%	2022

Datasets:

The Table 2 provides an overview of several widely recognized video datasets that have been used extensively in the research community for video analysis, particularly in the context of deep learning and computer vision. These datasets vary in terms of content, class diversity, and application focus, offering a range of challenges and opportunities for advancing video processing techniques. Each dataset is briefly described, highlighting the number of video clips, the nature of the classes, and the source from which the data was collected. The year of publication is also included to give context to the dataset's relevance and usage over time.

UCF-101 [19]: This dataset comprises 13,320 video clips across 101 different classes, providing a diverse collection of actions. The videos are sourced from YouTube, making it a popular choice for research in action recognition since its release in 2012.

HMDB51 [9]: Featuring at least 101 video clips per class across 50 classes, this dataset combines content from YouTube and movies. Released in 2011, it remains a key resource for studying human motion and behavior in various contexts.

ActivityNet [2]: Focused on human activity, this dataset includes 20,000 video clips spanning 200 classes. It is widely used for research on activity recognition, with data sourced from web searches. ActivityNet was introduced in 2015.

YouTube8M [5]: With an extensive collection of 800,000 video clips across 1,000 classes, this dataset represents a large-scale effort to categorize a wide array of video content. The videos are collected from YouTube video URLs, making it a critical resource for large-scale video analysis since its release in 2016.

UCSD [8]: This dataset is designed for binary classification, specifically for identifying the presence or absence of anomalies in video clips. It was manually curated and released in 2014, focusing on anomaly detection research.

UMN [14]: Another dataset aimed at binary classification, UMN is used to distinguish between normal and abnormal behavior. It was manually recorded using a head-mounted camera and has been a significant dataset for behavior analysis since its release in 2009.

Avenue [13]: This dataset targets abnormal event detection with 37 video clips, capturing various scenarios on the CUHK campus avenue. It has been a valuable resource for abnormality detection studies since 2013.

TABLE 2. Summary of the dataset

Ref.	Name	Description	Source	Year
[19]	UCF-101	The dataset has 13320 video clips with 101 different classes .	YouTube	2012
[9]	HMDB51	There are at least 101 video clips in each of the 50 classes in the dataset.	YouTube, Movies	2011
[2]	ActivityNet	The dataset is based on human activity and 200 classes with 20,000 video clips.	Web- search	2015
[5]	Youtube8M	The dataset has 1000 classes with 800000 video clips collected from YouTube video URLs.	YouTube	2016
[8]	UCSD	The dataset is for binary classification as an anomaly has or not.	Manually	2014
[14]	UMN	The dataset is for binary classification as normal and abnormal behavior	Manually by head-mounted camera	2009
[13]	Avenue	Abnormal event detection with 37 video clips	CUHK campus avenue	2013
[10]	UCF sports	The dataset is collected from various sports instances through 125 people to recognize the sports action.	Broadcast from BBC and ESPN	2013

UCF Sports [10]: Collected from various sports broadcasts on BBC and ESPN, this dataset includes videos of 125 individuals performing sports actions. It is used for recognizing and analyzing sports activities and has been available to researchers since 2013.

These datasets have played a crucial role in advancing research in video processing, providing benchmarks and challenging scenarios for developing and testing new algorithms in areas such as action recognition, anomaly detection, and behaviour analysis

3. ABNORMAL ACTIVITY DETECTION AND CHALLENGES

Abnormal activity identification in real-time videos is an important use of deep learning in computer vision, particularly in surveillance, security, and public safety. This work entails identifying behaviours or occurrences that depart from the norm within a video sequence, such as unexpected crowd movements, unauthorised access to restricted locations, or any other activity that could indicate a potential threat or abnormality. The idea is to detect these behaviours as they occur, allowing for rapid responses and interventions.

Several variables contribute to the difficulty of detecting aberrant activity:

Variability in Normal and Abnormal Behaviours: One of the most difficult difficulties is determining what constitutes "normal" and "abnormal" behaviour, which can differ dramatically between environments and contexts. Activities that are normal in one context may be regarded abnormal in another, making it challenging to create a one-size-fits-all paradigm. Furthermore, aberrant occurrences are frequently rare, creating an imbalanced dataset that hinders deep learning model training.

Real-Time Processing Requirements: To detect aberrant actions in real time, models must be able to process video input quickly and accurately. This requires processing enormous amounts of data at high speeds, which can be computationally intensive. Ensuring that the models are efficient enough to work in real time while maintaining accuracy is a huge problem, especially when dealing with complicated video data.

Noise and Variability in Video Quality: Video data is frequently affected by noise, motion blur, changing lighting conditions, and resolution variances, all of which can have an impact on detection model performance. Preprocessing techniques like as normalisation, noise reduction, and frame alignment are critical, but they must be used efficiently to avoid delays in real-time processing.

Occlusions and Ambiguities: In busy or complicated settings, objects and activities may be partially obscured, making it harder for the model to accurately recognise and classify behaviours. Furthermore, small or ambiguous movements may be difficult to classify as normal or abnormal, resulting in potential false positives or negatives.

Scalability and deployment: Implementing abnormal activity detection systems over large-scale monitoring networks raises scalability concerns. The models must be able to withstand changes in camera angle, field of view, and video quality across many sources. Furthermore, application in real-world scenarios necessitates consideration of hardware limits, network capacity, and integration with current security solutions.

Privacy Concerns: Real-time video monitoring, especially when dealing with aberrant behaviour detection, creates serious privacy concerns. Individual identification and activity tracking must be done with caution in order to protect personal privacy while maintaining security. Developing approaches that balance effective monitoring with privacy protection is a continuous problem.

To address these issues, contemporary deep learning research focusses on creating more complicated models capable of learning from imbalanced datasets, increasing the efficiency of real-time processing, and improving the durability of detection methods under changing conditions. The use of hybrid models, which combine Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures, shows promise in overcoming some of these challenges by leveraging each approach's strengths to improve the detection and classification of abnormal activities in real-time videos.

4. DISCUSSION & CONCLUSION

This article presents a thorough examination of deep learning techniques used for video processing in the field of computer vision. The scope of our investigation included a diverse array of preprocessing approaches, models, and datasets that have played a crucial role in advancing the discipline. In this review, we have discovered several important facts. One of these findings is the existence of extensive datasets that have greatly aided in the creation of strong video processing applications. Deep learning methods have demonstrated significant promise in improving pose identification, activity detection, video analysis, and general comprehension of videos. The initial research conducted in ImageNet and AlexNet for image processing has established the basis for these improvements, enabling the development of more intricate models designed specifically for video data. The study focusses on an extensive collection of research that revolves around using deep learning techniques for video processing. This includes tasks such as analysing behaviour, recognising human actions, and detecting anomalies in crowds. Nevertheless, despite the advancements achieved, there are still some obstacles that need to be overcome. Significant challenges persist in terms of the variability in video quality, temporal and motion-related problems, and the identification of anomalous activity. These issues encompass not only technical aspects but also incorporate intricacies associated with real-time processing, scalability, and privacy concerns.

Our study has found that despite the availability of improved methodologies and datasets, the constraints highlighted earlier still hinder the complete realisation of deep learning's potential in video processing. Tackling these obstacles is crucial for the ongoing advancement and ingenuity in this field. In the future, it is important to concentrate on addressing these challenges by creating and implementing sophisticated models. Hybrid models that combine Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures show potential in dealing with the intricacies of real-time video processing, enhancing the identification of abnormal activities, and handling the variability in video data. Furthermore, it is imperative to investigate novel techniques for optimising preprocessing and temporal encoding in order to improve the efficacy of deep learning models in video analysis.

In addition, it is important for future study to examine the ethical consequences of video processing, specifically with regards to privacy and the ethical utilisation of surveillance technologies. By prioritising these specific areas, the discipline may progress towards developing more efficient and dependable systems for practical applications, guaranteeing that video processing based on deep learning continues to advance and satisfy the increasing requirements of many industries.

REFERENCES

[1]. Basak, H., Kundu, R., Singh, P.K., Ijaz, M.F., Woźniak, M., Sarkar, R.: A unionof deep learning and swarm-based optimization for 3d human action recognition. Scientific Reports 12(1), 5494 (2022)

[2]. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: Alarge-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)

[3]. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and thekinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)

[4]. Elboushaki, A., Hannane, R., Afdel, K., Koutti, L.: Multid-cnn: A multidimensional feature learning approach based on deep convolutional networks for gesture recognition in rgb-d image sequences. Expert Systems with Applications 139, 112829 (2020)

[5]. Fernando, B., Gould, S.: Learning end-to-end video classification with rankpooling. In: International Conference on

Machine Learning. pp. 1187–1196. PMLR (2016)

[6]. Gong, S., Shi, Y., Jain, A.: Low quality video face recognition: Multi-mode aggregation recurrent network (marn). In: Proceedings of the IEEE/CVF InternationalConference on Computer Vision Workshops. pp. 0–0 (2019)

[7]. Islam, M.M., Bertasius, G.: Long movie clip classification with state-space videomodels. In: European Conference on Computer Vision. pp. 87–104. Springer (2022)

[8]. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Largescale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)

[9]. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large videodatabase for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)

[10]. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization incrowded scenes. IEEE transactions on pattern analysis and machine intelligence 36(1), 18–32 (2013)

[11]. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)

[12]. Logeshwaran, J., Ramkumar, M., Kiruthiga, T., Pravin, R.S.: Svpa-the segmentation based visual processing algorithm (svpa) for illustration enhancements in digital video processing (dvp). ICTACT Journal on Image and Video Processing 12(3), 2669–2673 (2022)

[13]. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision. pp. 2720–2727 (2013)

[14]. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using socialforce model. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 935–942. IEEE (2009)

[15]. Pashchenko, F.F., Amosov, O.S., Amosova, S.G., Ivanov, Y.S., Zhiganov, S.V.: Deep neural network method of recognizing the critical situations for transport systems by video images. Procedia Computer Science 151, 675–682 (2019)

[16]. Podder, K.K., Chowdhury, M.E., Tahir, A.M., Mahbub, Z.B., Khandakar, A., Hossain, M.S., Kadir, M.A.: Bangla sign language (bdsl) alphabets and numerals classification using a deep learning model. Sensors 22(2), 574 (2022)

[17]. Savakis, A., Shringarpure, A.M.: Semantic background estimation in video sequences. In: 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN). pp. 597–601. IEEE (2018)

[18]. Shukla, U., Mishra, A., Jasmine, S.G., Vaidehi, V., Ganesan, S.: A deep neuralnetwork framework for road side analysis and lane detection. Procedia Computer Science 165, 252–258 (2019)

[19]. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of human actions classesfrom videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

[20]. Vrskova, R., Hudec, R., Kamencay, P., Sykora, P.: Human activity classificationusing the 3dcnn architecture. Applied Sciences 12(2), 931 (2022)

[21]. Wang, X., Miao, Z., Zhang, R., Hao, S.: I3d-lstm: A new model for human actionrecognition. In: IOP Conference Series: Materials Science and Engineering. vol. 569, p. 032035. IOP Publishing (2019)

[22]. Wu, G., Song, S., Li, L.: Video summarization generation model based on transformer and deep reinforcement learning. In: 2023 8th International Conference on Computer and Communication Systems (ICCCS). pp. 916– 921. IEEE (2023)

[23]. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative cnn video representation forevent detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1798–1807 (2015)

[24]. Yousaf, K., Nawaz, T.: A deep learning-based approach for inappropriate contentdetection and classification of youtube videos. IEEE Access 10, 16283–16298 (2022)

[25]. Khalifeh, I., Murn, L., Mrak, M., Izquierdo, E.: Efficient convolution andtransformer-based network for video frame interpolation. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1050–1054. IEEE (2023)

[26]. Amosov, O., Amosova, S., Ivanov, Y., Zhiganov, S.: Using the ensemble of deepneural networks for normal and abnormal situations detection and recognition in in the continuous video stream of the security system. Procedia computer science 150, 532–539 (2019)

[27]. Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.