Video Content Analysis Using Deep Learning Methods

Gara Kiran Kumar[1(✉)] and Athota Kavitha[2]

[1] Department of Computer Science and Engineering, Anurag University, Hyderabad, India
kirankumarcse@anurag.edu.in

[2] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Kukatpally, Hyderabad, India

Abstract. With the emergence of low-cost video recording devices, the internet is flooded with videos. However, most videos are uncategorized, necessitating video content analysis. This review effort addresses visual big data feature extraction, video segmentation, classification, and abstract video challenges. Exploring compressive sensing, deep learning (DL), and kernel methods for various tasks in video content analysis include video classification, clustering, dimension reduction, event detection, and activity recognition. DL is used to examine video footage recognition and classification. This study examines the algorithms' flaws and benefits when applied to datasets. The classification approaches used Naive Bayes, support vector machine (SVM), and Deep Convolution Neural Network (DCNN) with Deer Hunting Optimization (DHO). Other approaches have higher false discovery and alarm rates than the DCNNDHO algorithm.

Keywords: video content segmentation · video content recognition · classification · Deep Convolution Neural Network (DCNN) with Deer Hunting Optimization (DHO)

1 Introduction

In multimedia, a combination of videos and images, generation by individuals, the boom is caused by integrated cameras and handheld devices availability with a large storage capacity complemented using the cloud. Through fast broadband connections and high bandwidth, these contents are shared. Social media's reaching power assists this. By 2025, it was estimated that there were approximately 9 trillion images stored on cameras, storage media, and in the cloud around the globe and that 2.4 trillion additional digital pictures will be taken in 2025 alone. In July 2015, it was reported that more than 400 h of Video were posted to YouTube every minute [1], continuing a similar pattern. The growing deployment of Closed Circuit Television (CCTV) cameras, which capture large quantities of media in public and private spaces to improve surveillance and deter crimes, adds to the phenomenon. Because of substantial real-life applications, a significant research area of interest is video content analysis, which is used to perform its functionalities. Video streams are analyzed automatically using Video Content Analysis

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 471, pp. 222–243, 2023.
https://doi.org/10.1007/978-3-031-35081-8_18

(VCA). Spatial and temporal events are determined and detected using this. About video content, helpful information is provided using this [2]. For analyzing video contents captured using surveillance systems, this is typically inherent. VCA has been used in a broad range of applications such as traffic control, health care, transportation, detecting intruders, and counting the number of people incoming or leaving locations. With VCA, the monitoring system is improved. It can accurately trace all objects' movements, where they came from, where they went, and when. This work is used to reduce the amount of time spent on video search while also improving the accuracy of the results.

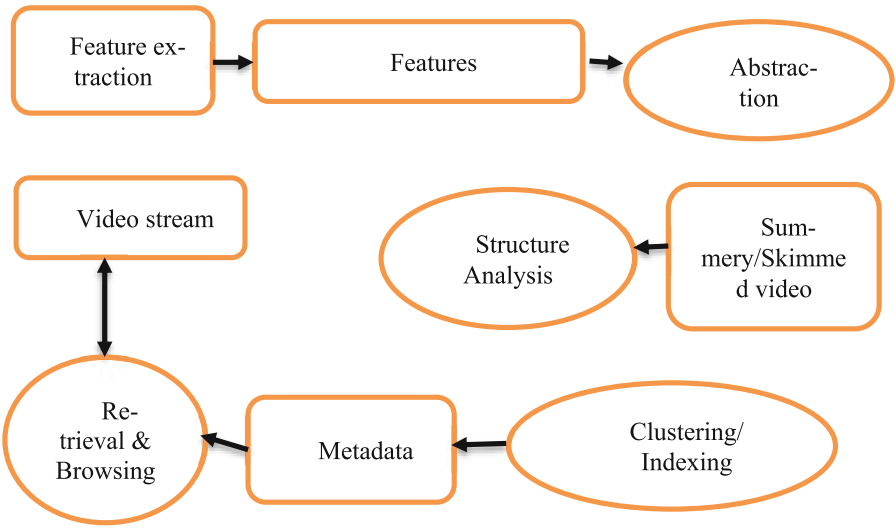

Fig. 1. Flowchart of video content analysis.

In a video program, a significant information source is a visual content. In video-content analysis, from multimedia sources, attributes are extracted using an effective strategy [3]. For professional and consumer applications, the video program is characterized by using text, audio, and video components cooperative and combined analysis. Feature extraction is critical in video indexing, as shown in Fig. 1. In content representation, attributes effectiveness is based on indexing schemes' effectiveness. However, mapping extracted features like motion, layout, structure, Shape, Texture, and color into semantics like car-racing, people, and outdoor and indoor scenes is not done quickly. Term video segmentation refers to decaying video data into meaningful elementary parts, which strongly correlates with the real world in video data. The video segmentation result is a segment set that collectively covers the actual entire video data. The significant difference between an image signal and a video signal is that a video signal consists of temporal information, which includes camera motion and introduces the concept of object motion. Therefore Video has temporal nature as well as spatial nature. Segmentation of Video can thus be temporal, spatial, or Spatio-temporal [4]. In the spatial domain, frame segmentation is like a static image. The video frame's sequence

segmentation is termed shot detection or temporal segmentation in the temporal domain (Fig. 2).

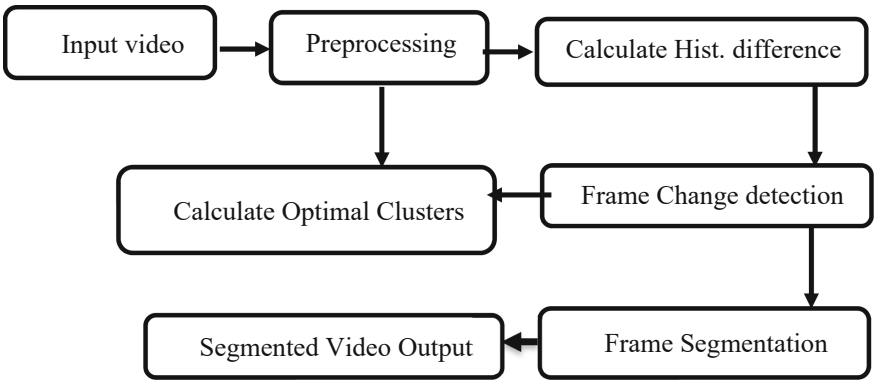

Fig. 2. Block diagram of Video Segmentation.

This classification of video segmentation is based on the features available with the input video data. Under this classification, it still has many segmentation techniques based on what type of feature the method is working for segmenting the Video. They are techniques based on semantics, object, content, Edge, Region, pixel, etc. Extraction of low-level features is done at first, and these features are mapped to category or concept labels by training and employing specific classifiers. For video genre identification, the bag-of-visual-words (BoW) model with k-nearest neighbor (KNN) classifiers is adopted in [5]. In this technique, for characterizing video sequence's every frame, a technique based on unsupervised probabilistic latent semantic analysis is used, and it produces a generic framework to analyze sports videos. Frames are classified into any of the following four groups: outer-field, long, mid and close-up view. In mobile video applications, to conform to sports type, four auxiliary datasets are used in the multimodal technique, which includes audio data, spatiotemporal and spatial visual information, and sensor-embedded mobile devices [6]. In infusion models, all possible modality combinations are integrated using auxiliary input data for classifying video types. Numerous research and methodologies are introduced, but the video content classification dataset accuracy is not ensured significantly. Therefore, this survey study suggests feature extraction, video segmentation, and video classification algorithms. So, this comprehensive survey focused on classifying video content dataset performance by using efficient and effective methods. The present study is done in the state of the art of various video content recognition and classification methods.

This study is organized in 4 sections. The literature review is presented in Sect. 2, the experimental results are provided in Sect. 3, and the conclusion of this survey study is given in Sect. 4.

2 Related Work

Various authors present different approaches for video classification, video segmentation, and feature extraction. In recent times, current and most often used techniques are discussed in this section.

2.1 Review on Feature Extraction Techniques for Video Content Dataset

In content-based visual information retrieval (CBVIR), the typical visual feature extraction applications set is parallelized and optimized by Chen et al. (2007) in [7]. Automatic video management is a mass-market application due to explosive growth in video data. One of the best solutions for this CBVIR. Low-level feature extraction forms the base for this CBVIR system. The MPEG-7 standard is used for feature extraction for high-level semantic concept indexing. A highly computation-intensive task is CBVIR, and in this CBVIR, the highly time-consuming component is the extraction of low-level visual features. In recent data, available multi-core processors' computing power is utilized fully for accelerating CBVIR. This is due to the recent advancements in a multi-core processors. Underlying parallel and optimization methods are examples of video analysis applications, and on multi-core systems, they are used in other applications for enhancing performance. These parallel applications' detailed performance analysis is conducted on a dual-socket, quad-core system. Possible bottleneck causes are identified using this analysis, and for scalability enhancement, this suggests avenues. In realtime performance, applications are made highly powerful in terms of accuracy. In broadcasted sports videos, the Superimposed Caption Box (SCB) interpretation is illustrated by Shih et al. (2008) in [8]. In those videos, the SCB template is not presumably given as s priori. Video content's digested vital information is represented using embedded captions in sports video programs. Known character bitmaps and SCB templates are assumed in most previous studies. This paper's significant contributions include identifying and extracting caption templates, modeling and extracting symbols, and identifying symbols and captions semantic interpretation. SCB contents understanding is done using this algorithm for various commercial sports video programs, as shown in experimental results. To rectify Co-occurrence Histograms of Oriented Gradients (CoHOG) drawbacks, a technique is suggested by Pang et al. (2012) in [9]. With respect to detection accuracy and computation, better results are produced using this technique. Significant contributions of this technique include gradient magnitude information used in addition to orientation information for enhancing detection accuracy significantly. A novel gradient decomposition technique is used, which is regarded as force's decomposition, and a combined approach is used, which is considered force's composite. Deep learning models can automatically learn the optimal features without human interaction. For example, digit identification, image classification, feature learning, visual recognition, musical signal processing, and NLP [10] can benefit from the deep learning framework's robust and accurate feature learning in supervised and unsupervised scenarios. Convolutional neural networks (CNNs) were used to extract visual features for sentiment analysis applications because of their recent success in feature extraction. Using CNN, [11] created a novel visual sentiment prediction system. The method uses transfer learning from a CNN pre-trained on large-scale data for object identification to predict

sentiment. The suggested system eliminates the need for domain knowledge for visual sentiment prediction. In 2014, You et al. [12] used 2D-CNN for visual sentiment analysis, fine-tuning deep learning networks to filter out noisy training data. They employed domain transfer learning to boost performance. [13] suggested a deep 3D convolutional network for spatiotemporal feature extraction in 2015. A softmax output layer and eight convolution layers are present in this network. The network has shown excellent spatiotemporal feature extraction. Poria et al. [14] proposed using a convolutional recurrent neural network to extract visual characteristics from multimodal sentiment analysis and emotion identification datasets.

Under the two-stage convolution framework, obtained CoHOG features for suppressing the aliasing effect in a great manner. Usage of integral histogram in two-stage convolution framework enhances feature extraction process speed. A mean spatial filter shows better performance than a linear spatial filter. In the positive class, Incremental Principal Component Analysis (IPCA) (PCA) is employed to minimize CoHOG feature's dimensionality. For key-frame extraction, a framework for modeling semantic contexts is presented by Yong et al. (2013) in [15]. Extracted video frame's semantic context. Monitored its sequential changes and located its significant novelties using a one-class classifier. Working with wildlife video frames, for representing semantic context in scene, image segmentation is done in framework. Then features are extracted, and image blocks are matched. The semantic label's co-occurrence matrix is constructed. Better key-frame extraction is done using high-level semantic modeling, as shown in experimental results compared with low-level features usage. For action recognition, an effective feature extraction technique is developed by Kantorov & Laptev (2014) in [16]. Over-current years, action identification's precision has been enhanced incessantly. Characteristic extraction's low speed and succeeding recognition thwart present techniques from scaling up to a real-size crisis. It concentrates on the initially developed extremely proficient video characteristics through movement information in video compression. Subsequently exploring feature encoding by Fisher vectors and reveal precise action identification with fast linear classifiers. This technique progresses the video feature mining speed, feature encoding speed and action categorization speed by two orders of magnitude with minor reduction in recognition accurateness. Primarily, to evade the long latency owing to frame level calculation, it accepts the ayer parallel restructured box kernel to reinstate iterated Gaussian blur operations. A cascade classifier ensembles bailing-based reliable vehicle classification technique is provided by Zhang et al. (2014) in [17]. For image description, two feature extraction methods are introduced in this designed system. Namely, Pyramid Histogram of Oriented Gradients (PHOG) and Gabor wavelet transform. Image characteristics are hauled out for various vision tasks by implementing Gabor transform. In added discriminating information explanation, PHOG has its advantages. With the refused option, cascade classifier ensembles anticipate an extremely consistent classification technique for holding the state of affairs. If there is sufficient ambiguity, no conclusion is made. The primary ensemble exhibits heterogeneous nature, and it has various classifiers in addition to random forest, Supp, SVMs, Multiple-Layer Perceptrons (MLPs), and CNNs. A second classifier ensemble is used to enhance classification reliability further. This has a base MLPs coordinated

using an ensemble Meta-learning method called Rotation Forest (RF). For both ensembles, the rejection option is proficient in concerning consensus degree from bulk voting to confidence measure and nonparticipation for categorizing indefinite samples if consensus degree is inferior to the threshold. For extracting semantic events from sports webcast text, an unsupervised technique is proposed by Chen et al. (2014) in [18]. First, in webcast text descriptions, filtering of unrelated words is done. The significant event classes are formed by clustering these filtered descriptions. At last, for every event class, keywords are extracted. As shown in experimental results, significant event texts are extracted using this technique. Video summarization and indexing are done using these event texts. Further, for text event retrieval, a hierarchical searching technique is given by this. In a large video database, automated video search and video indexing are implemented by Chivadshetti et al. (2015) in [19]. Personalized results are also presented. There are three various phases in this system. Key-frame detection and video segmentation are performed in the first phase to extract meaningful features. Over key-frames, ASR, HOG, and OCR algorithms are applied for extracting textual keywords. Edge, Texture, and color features are extracted in the third phase. All extracted features are stored in the database, and then classification is done concerning query video. At last, on extracted features, performed the search similarity measure. As per the user's interest, personalized re-ranking results with outputs are presented. Between Local Binary Patterns (LBPs) coded frames, based on correlation coefficient quotients, an effective technique is suggested by Zhang et al. (2015) in [20]. This technique has two parts: abnormal point detection and feature extraction. Every video frame is coded using LBP in feature extraction. Then, the correlation coefficient's quotients are computed among sequential LBP-coded frames. Chebyshev inequality is used twice for deleting and inserting localization in abnormal point detection. After this inequality application, decision thresholding is used to detect abnormal points. High detection accuracy with low computational complexity is achieved using this technique, as shown in experimental results. For facial expression analysis, a novel framework is introduced in video sequences by Zhao et al. (2017) in [21], where static and dynamic information is used. For locating facial points, breaking facial regions from the background, and correcting in-plane head rotation, adapted an incremental formulation based on discriminative deformable face alignment. Then, extracted spatial-temporal motion Local Binary Pattern (LBP) feature and for producing descriptors, with Gabor multi orientation fusion histogram, these features are integrated. Facial expressions and dynamic and static texture information are reflected using this. At last, facial expressions are classified by applying a multiclass SVM classifier based on a one-versus-one strategy. Methods using single descriptors are outperformed by this integrated framework, as illustrated in experimentation results. Cohn-Kanade (CK) + facial expression dataset is used in experimentation. On Oulu-CASIA VIS, MMI, and CK + datasets, better performance is shown by this technique when compared with other state-of-the-art techniques. From video frames or images, for computing transform coefficients (features), a novel technique is introduced by Abdulhussain et al. (2019) in [22]. Video frames and the image's local visual contents are represented using these features. Using a standard imaging technique, traditional feature extraction methods are compared. Further, for detecting transitions in Shot Boundary Detection (SBD) applications, video frames employed a fast feature technique. The SBD

application's performance is evaluated using standard TRECVID 2005, 2006, and 2007 video datasets. Compared with traditional techniques, the computational cost is reduced significantly using this algorithm (Table 1).

Table 1. Inference of existing feature extraction methods.

S.No.	Method	Advantages	Disadvantages
1	Accelerating Video Feature Extraction method [12]	• It discards irrelevant attributes from the given dataset • It is used for low-level visual feature extractions significantly	• It increases the overlapping of classes and can introduce additional noise
2	CoHOG feature extraction method [14]	• Computational complexity is reduced significantly, and essential features are extracted • It provides higher accuracy	• For meaningful video content, classification performance is low
3	Wildlife video key-frame extraction method [15]	• High-level semantic modeling achieves better key-frame extraction	• In a few cases, it has an issue with computational complexity
4	KNN, MLP, and SVM algorithms [17]	• Faster calculation time for implementation • Relatively memory efficient	• In a few cases, system efficiency is reduced
5	An unsupervised approach [18]	• It provides higher accuracy, sensitivity, specificity, and precision	• However, it has a problem with expensive rates
6	Integrated feature extraction and personalization method [19]	• It is used for re-ranking results as per interest is presented to the users	• In a few cases, video content classification precision result is lower
7	LBP method [20]	• The method has high detection accuracy and low computational complexity	• But it has an issue with noise rates
8	LBP and SVM [21]	• It is used to provide more accurate classification results	• However, it has an issue with long training time
9	Fast feature extraction algorithm [22]	• This algorithm is used to efficiently represent the local visual content of images and video frames • It reduces the computational cost	• It requires higher memory

2.2 Reviews of Video Segmentation Methods

A segmentation technique based on the Genetic Algorithm (GA) is suggested by Kim et al. (2006) in [23]. Moving objects are tracked as well as extracted using this technique automatically. Temporal and spatial segmentation is done mainly in this. Every frame is split as a region with accurate boundaries is formed using spatial segmentation, and every frame is split as a foreground and background area using temporal segmentation. Individuals evolved using Distributed Genetic Algorithms (DGAs) are used for performing spatial segmentation. From the previous frame's segmentation results, initiated the individuals. This is not the case in standard DGAs. Thus mating operator is used for evolving unstable individuals of actual moving objects. According to intensity difference, adaptive thresholding is done between two consecutive frames for temporal segmentation. For object extraction, results of temporal and spatial segmentation are combined. Natural correspondence established using spatial segmentation technique is used for performing the track. A novel technique for segmenting scenes automatically and representing them semantically is introduced by Zhu et al. (2009) in [24]. At first, a rough-to-fine algorithm is used for detecting video shots. Then, within every shot, adaptively selected key-frames with hybrid features and template matching are used to remove redundant key-frames. In the third stage, according to the visual similarity between shot activities and video content's temporal constraint, the same scene is formed by clustering Spatio-temporal coherent shots. At last, scene content is represented semantically on video retrieval under a typical character's full analysis of continuously recorded videos for satisfying human demand. On different TV programs and film genres, performed this algorithm. Interesting video content is retrieved effectively using this method, as shown in experimentation results. Tree-structured graphical models based on occlusion aware semi-supervised video segmentation algorithm are implemented by Budvytis et al. (2011) in [25]. This algorithm is an implementation-friendly one. Pixel labels with their uncertainty estimation are delivered using this algorithm. Supervision is employed for tackling task-specific segmentation problems, where the user predefines the semantic objects. Patch-based undirected mixture model's tree-structured approximation-based video model is used in this problem. A soft label Random Forest classifier and novel time series are included in this feedback. In complex and lengthy road scene sequences, multiclass segmentation problems, and cutting out foreground objects, the model's efficiency is demonstrated using this. There are several applications of this result. For discriminative training models, labeled video data is harvested using these results. For video segmentation, priors are developed using large-scale statistical analysis and articulation/pose/shape learning. A Parametric Graph Partitioning (PGP) based robust as well as effective video segmentation framework is used by Yu et al. (2015) in [26]. Nodes are identified and removed between cluster edges using this fast and parameter-free graph partitioning technique, and it forms a node cluster. Without pre-specified bandwidth parameters and cluster number, Spatio-temporal volume's clustering is performed using PGP other than its computational efficiency. Video segmentation is made highly practical to be used in applications. Sub-volumes are processed using this PGP framework. Performance is enhanced further because of this. In other streaming video segmentation techniques, performance is degraded by sub-volume processing. Chen Xiph.org and SegTrack v2 datasets are used for evaluating this PGP technique. In 3D segmentation

running time and metrics, related state-of-the-art algorithms are outperformed. Using deep learning convolutional neural nets, a three-dimensional video segmentation technique is introduced by Piramanayagam et al. (2020) in [27]. For generating initial pixel groups, global boundary maps computed using deep learning techniques are used in addition to the computed local gradient at every pixel location. Traversal to high from low gradient region is done for the same. These initial pixel groups are refined using the local clustering technique. In Video's homogeneous regions, refined sub-volumes are chosen as initial seeds, and according to intensity similarity, they are combined iteratively with adjacent groups. A video's color boundaries terminated volume growth. A multivariate technique is used to merge the over segments obtained in the above steps hierarchically. For every frame, a final segmentation map is produced using this. On a quantitative and qualitative level, concerning computational efficiency and segmentation quality, favorable results are produced using the proposed methodology, as shown in experimental results. A video segmentation benchmark dataset is used in experimentation (Table 2).

Table 2. Inferences on existing video segmentation techniques.

S.No.	Methods	Merits	Demerits
1	Distribute Genetic Algorithm (DGA)-based segmentation method [23]	• This algorithm is used to reduce the search time • It reduces the computational cost	• In a few cases, the segmentation accuracy
2	Video scene segmentation and semantic representation method [24]	• It is used for efficient retrieval of exciting video content • It provides higher segmentation accuracy	• In a few cases, redundant keyframes are an issue
3	Semi-supervised video segmentation algorithm [25]	• It includes harvesting labeled video data for training discriminative models, shape or pose or articulation learning, and large scale statistical analysis to develop priors for video segmentation	• In a few cases, error rates are a problem
4	Parametric Graph Partitioning (PGP) [26]	• It increases the video segmentation performance effectively • Running time is fast	• Expensive is an issue
5	Deep learning convolutional neural nets [27]	• It provides higher quality and computational efficiency	• However, it has a problem with lower resolution

2.3 Reviews of Video Content Classification Algorithms

For online-video sharing Web sites content classification, a text-based framework is suggested by Huang et al. (2010) in [28].Various user-generated data types called comments, descriptions, and titles are used as proxies for online videos. Extracted three text features, namely content-specific, syntactic and lexical features. Three classification methods based on features, namely, Support Vector Machine, Naïve Bayes, and C4.5, are used for video classification. From candidate videos, user-generated data identified by using user-given keywords on YouTube are collected first for evaluating this framework. Then, experiment data is formed by random selection of collected subset data, and they are tagged manually by users. With around 87.2% accuracy rate, based on users' interests, online videos are classified using this technique are shown in experimental results, and videos are discriminated using all three text feature types. In experimentation, Naïve Bayes and C4.5 methods are outperformed by SVM. On video-sharing Websites, accurate video-classification results are beneficial, as demonstrated in results to identify implicit cyber communities. Fan et al. (2016) presented a video-based emotion recognition system in [29] and submitted it to EmotiW 2016 Challenge. In a late-fusion fashion, 3D Convolutional Networks (C3D) and Recurrent Neural networks (RNN) are combined to form a hybrid network, a core module in this system. Variously, motion and appearance information is encoded using C3D and RNN. Specifically, a convolutional neural network (CNN) extracts appearance features over individual video frames and is given to RNN as input, and motion features are encoded later. Video motion and appearance are modeled simultaneously using C3D.In the training set, without using any additional emotion-labeled video clips, around 59.02% accuracy is achieved using this system with the audio module, whereas, EmotiW 2015 winner achieves 53.8% accuracy. As shown in experimentation results, video-based emotion recognition is enhanced significantly by combining C3D and RNN. A novel action recognition technique is presented by Ullah et al. (2017) in [30], where deep bidirectional LSTM (DB-LSTM) and convolutional neural network (CNN) is used for processing video data. From videos, every sixth frame extracts deep features and is used for minimizing complexity and redundancy. Then, the DB-LSTM network is used for learning sequential information available in frame features. For increasing depth, in the backward and forward pass, multiple layers are stacked together in DB-LSTM. Long-term sequences are learned using this technique, and for a certain time interval, features are analyzed for processing lengthy videos. On three benchmark datasets, namely, HMDB51, YouTube 11 Actions, and UCF-10, significant action recognition enhancement is shown using this technique, as illustrated in experimental results compared with other state-of-the-art action recognition techniques. For a crowd, density level classification, violent behavior detection, and simultaneous crowd counting, ResnetCrowd, a deep residual architecture, is presented by Marsden et al. (2017) in [31]. Constructed a new dataset with 100 images, and it is termed as Multi-Task Crowd for evaluating and training multi-objective techniques. For crowd density level classification, violent behavior detection, and simultaneous crowd counting, this new dataset is used as the first computer vision dataset, which is fully annotated. As shown in experimental results, individual task performance is boosted for

all tasks using a multi-task approach. In ROC, AUC (Area under the curve), around 90% enhancement is achieved in violent behavior detection. On various other benchmarks, this trained Resnet Crowd model was evaluated, which highlights the crowd analysis model's superior generalization and is trained for multiple objectives. Video content analysis is discussed by Aljarrah et al. (2018) in [32]. This is focused on the rapid increase in real-life applications that use video content analysis to perform its functionalities. It is useful to review the videos recorded using video surveillance systems. An automated content analysis system is used instead of rewinding recorded videos over hours. A searchable text file with video content summarization is produced using this system. Object classification is used in this work for analyzing video surveillance content. A convolutional neural network model is used to detect and classify objects in a video. A text file with detected object classes and appearance time is generated for future usage. Only (I) frames are processed to speed up this heavy computational process. For emotional Big Data, a deep learning-based emotion recognition system is discussed by Hossain et al. (2019) in [33]. There is a video and speech in big data. In this system, for obtaining Mel-spectrogram, the speech signal processing in the frequency domain is done first. This Mel-spectrogram is assumed as an image. Then, a convolutional neural network (CNN) is given with this Mel-spectrogram. From a video segment, extracted some representative frames are fed to CNN in case of video signals. Two consecutive extreme learning machines (ELMs) are used for fusing two CNNs outputs. A support vector machine (SVM) is given for emotion's final classification with this fusion output. Two audio-visual emotional databases are used in this system evaluation. A system with ELMs and CNN's effectiveness is confirmed using experimental results. In videos, crowd event classification is discussed by Shri et al. (2019) in [34]. In computer vision-based systems, it is a highly challenging and important task. Huge video events are recognized using this crowded event classification system. In event classification, a difficult task is the model's decisive. This event classification model shows generalization capability in works with a large video count. In video classification, distinguishable features and powerful portrayals are derived using Deep learning's embodiment. From raw data, the extracted event features using a large number of videos with efficient and effective detection. A power classification model is provided using Convolutional Neural Network (CNN) for event recognition problems. From YouTube, I collected a high-quality 3000 frames for forming a new dataset. This includes four crowd event classes: shopping mall, Jallikattu, cricket, and marriage. Two Deep CNN infrastructures called VGG16 and baseline are used in this system for providing temporal evidence and detecting predefined events. In Video, centrality events are detected, and input video frames are tested automatically using the CNN model. Video event features are extracted using CNN from video input frames, and correct distinguishing of events is done. When compared with other models, around 100% enhanced results are provided using this system. For field sports videos, AlexNet Convolutional Neural Networks (AlexNet CNN) based effective shot classification technique is suggested by Minhas et al. (2019) in [35]. This network has an eight-layer structure, including three fully-connected layers and five convolutional layers, classifying shots as an out-of-field, close-up, medium, and long shots. Over a soccer and cricket video's diverse dataset, evaluated overall validation and training performance. These performances are boosted through feature maps dropout

layers and response normalization. Around 94.07% maximum accuracy can be achieved using this model when compared with standard Convolution Neural Network (CNN), K-Nearest Neighbors (KNN), Extreme Learning Machine (ELM), and Support Vector Machine (SVM). To prove this approach's superiority, this method is compared with baseline state-of-the-art shot classification techniques. Khan et al. (2019) developed an affordable in-vehicle snow detection system [36]. In real time, trajectory-level weather information is provided using this system. An SHRP2 Naturalistic Driving Study video data is used in this system, and machine learning approaches are also used. Three classification algorithms, namely random forest (RF), k-nearest neighbor (K-NN), and support vector machine (SVM), two image features based on Texture, namely local binary pattern (LBP), gray level co-occurrence matrix (GLCM), are used for training this snow detection models. An image dataset with three weather conditions, namely heavy snow, light snow, and clear, is used in the analysis. Around 86% of overall prediction accuracy is produced by GLCM features-based models, whereas around 96% of overall prediction accuracy is produced by LBP-based models, which is a greater one. In this study, a cost-effective snow detection system is used, and there is no need to have huge technical support. A single video camera is required for this. In real time, roadway weather conditions are detected with reasonable accuracy using simple mobile apps with a proper data connection. This is due to the advancements made in smartphone cameras. A movie trailer classification based on home action is proposed by Shambharkar et al. (2020) in [37]. This technique uses an optimized deep convolutional neural network in video sequences. At first, images are converted into their grayscale range. Pre-processing is done using adaptive median filtering. From video frames, the background is subtracted using the segmentation technique based on the threshold value, and the foreground portion is extracted. The segmented portion extracts motion features and visual features like Texture and color in the feature extraction stage. At last, for human action classifications, an optimized deep convolutional neural network (DCNN) is used for classifying mined features. DCNN weight values are optimized using introduced deer hunting optimization (DHO). In MATLAB environment, executed the human action classification based on DCNNDHO human action. The experimental results and comparisons are made between them, including the metrics like false discovery rate, F-measure, precision, specificity, sensitivity, false alarm rate, and accuracy. With and without the filtering process also, results are compared (Table 3).

Table 3. Inferences on existing video content recognition and classification methods.

S.No.	Methods	Merits	Demerits
1	C4.5, Naïve Bayes, and SVM methods [28]	• This algorithm demonstrated that accurate video-classification results are beneficial for identifying implicit cyber communities on video-sharing Web sites	• In a few cases, it is still too low for such an essential action recognition task
2	RNN with the C3D method [29]	• It is used to produce higher storage space with better efficiency for larger videos	• This method achieved a lower recognition accuracy
3	CNN and DB-LSTM method [30]	• This method is capable of learning long term sequences and can process lengthy videos by analyzing features for a certain time interval	• It has a problem with extracting low-level features
4	ResnetCrowd architecture [31]	• It increases the video classification accuracy	• Expensive is an issue
5	Convolutional Neural Network (CNN) model [32]	• It is used to speed up this heavy computational process	• However, it has a problem with video classification on a complex background
6	CNN, ELM, and SVM [33]	• It is used for the better emotion recognition system	• It has an issue with high computational power and a considerable volume of data
7	Deep learning-based CNN [34]	• It provides higher quality videos for crowd events	• It may suffer from the noisy Variation
8	SVM, KNN, ELM, and CNN [35]	• This model achieves the maximum accuracy	• Computational complexity is still an issue
9	GLCM, LBP, RF, SVM and KNN [36]	• It does not require much technical support and only needs a single video camera • It can effectively be used to detect roadway weather conditions in real time with reasonable accuracy	• It is computationally expensive

(*continued*)

Table 3. (*continued*)

S.No.	Methods	Merits	Demerits
10	DCNNDHO algorithm [37]	• It achieves higher accuracy, precision, and f-measure values	• However, it has an issue with background noise rates

The deep learning methods will be used for many applications like human action recognition and image retrieval [38–45].

3 Experimental Results

In MATLAB 2022a platform, employed an algorithms like CNN, RNN with C3D and DCNNDHO and studied the outcomes. Existing works are used for making comparison. With respect to false discovery rate, F-measure, precision, specificity, sensitivity, false alarm rate and accuracy, performance is evaluated.

Dataset 1:
UCF101 dataset is an extension of UCF50 and consists of 13,320 video clips, which are classified into 101 categories. These 101 categories can be classified into 5 types (Body motion, Human-human interactions, Human-object interactions, Playing musical instruments and Sports). The total length of these video clips is over 27 h. All the videos are collected from YouTube and have a fixed frame rate of 25 FPS with the resolution of 320 × 240.

Dataset 2:
The **Kinetics** dataset is a large-scale, high-quality dataset for human action recognition in videos. The dataset consists of around 500,000 video clips covering 600 human action classes with at least 600 video clips for each action class. Each video clip lasts around 10 s and is labeled with a single action class. The videos are collected from YouTube.

3.1 Accuracy

Model's overall correctness is determined using accuracy and it is a ratio between total actual classification parameters ($T_p + T_n$) to sum of all classification parameters ($T_p + T_n + F_p + F_n$). Accuracy is expressed as,

$$\text{Accuracy} = \frac{T_p + T_n}{\left(T_p + T_n + F_p + F_n\right)} \tag{1}$$

Accuracy metric evaluation and its comparison is illustrated in above shown Fig. 3. In that, various techniques are represented in x-axis and accuracy value is expressed in y-axis. Available methods are such as RNN with C3D and CNN algorithms provide lower accuracy whereas DCNNDHO algorithm provides higher accuracy for the UCF and kinetics datasets. Thus the result concludes that the DCNNDHO increases video classification accuracy for both datasets.

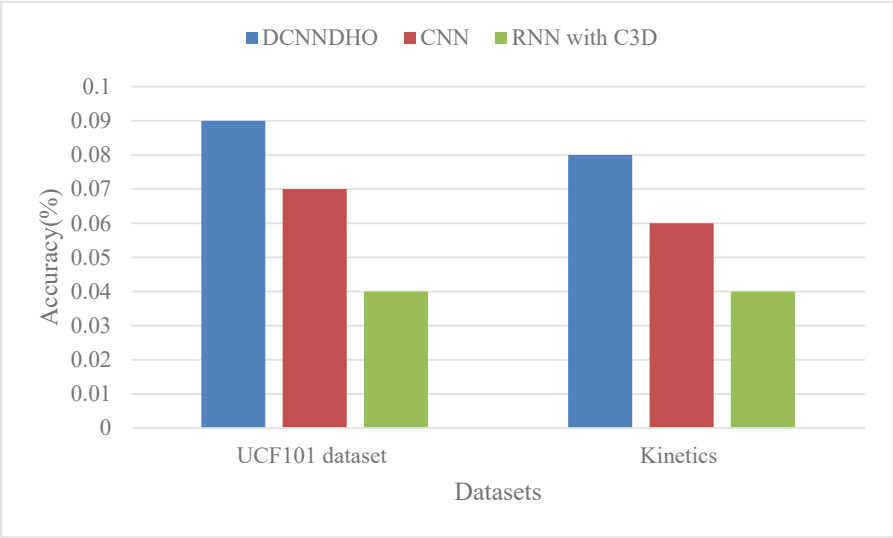

Fig. 3. Accuracy.

3.2 Precision

Precision value is computed as:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \tag{2}$$

Quality or accuracy is computed using precision value and quantity or completeness is measured using recall value. Computation of highly relevant results than irrelevant results by an algorithm is indicated using high precision value. For a class, ratio between true positives count to total elements which are labelled as positive class's count defines precision value in classification task.

Precision metric evaluation and its comparison is illustrated in above shown Fig. 4. In that, various techniques are represented in x-axis and precision value is expressed in y-axis. DCNNDHO algorithm provides higher precision while other existing RNN with C3D and CNN algorithms provide lower precision values for the given UCF and kinetics datasets. Thus the result concludes that the DCNNDHO algorithm increase the video classification accuracy for the both datasets.

3.3 Sensitivity

Actual positive's proportion which are identified correctly defines sensitivity value. This is also termed as recall or true positive rate. In some fields, it is termed as detection probability. It is also stated as a sick people's percentage who are identified correctly as with condition.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

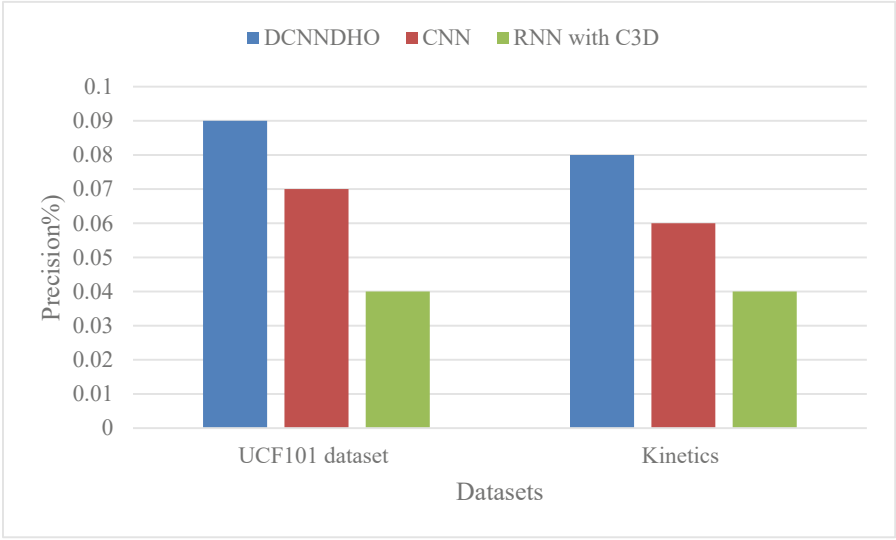

Fig. 4. Precision

where, True positive value is represented as TP and False Negative value is represented as FN.

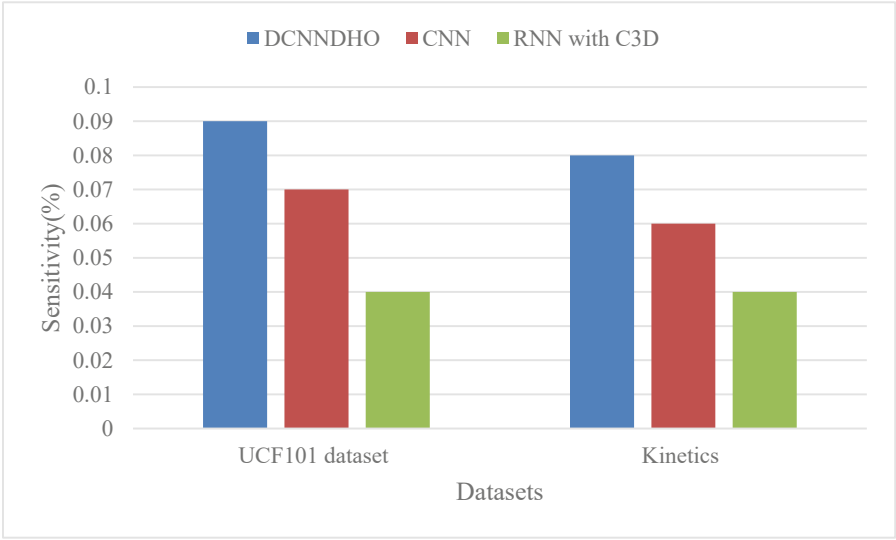

Fig. 5. Sensitivity.

Sensitivity metric evaluation and its comparison is illustrated in above shown Fig. 5. In that, various techniques are represented in x-axis and sensitivity value is expressed

in y-axis. Available methods are such as RNN with C3D and CNN algorithms provide lower sensitivity whereas DCNNDHO algorithm provides higher sensitivity for the UCF and kinetics datasets. Thus the result concludes that the DCNNDHO increase the video classification accuracy for the both datasets.

3.4 Specificity

Actual negatives proportion which are identified correctly as such gives specificity values and it is also termed as true negative rate. It is also stated as, healthy people's percentage who are identified correctly as not with condition.

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

where, True Negative value is represented as TN and False Positive value is represented as FP.

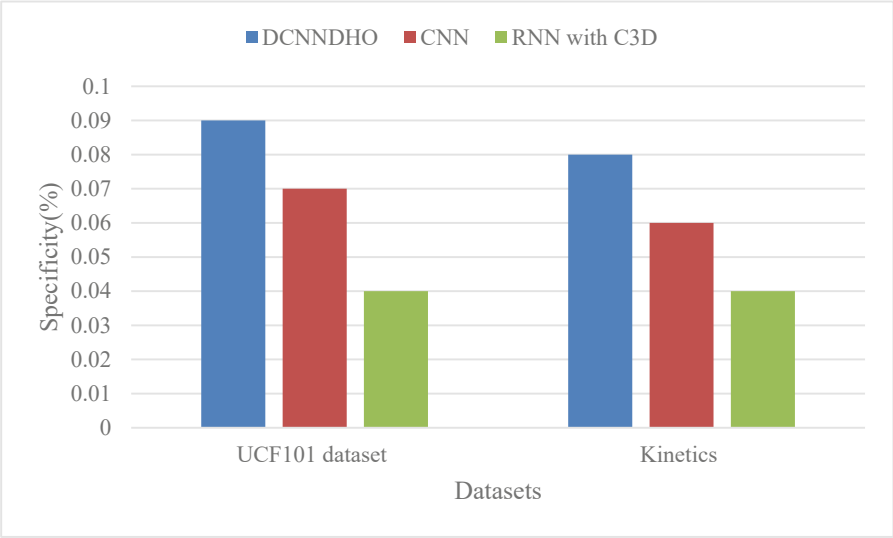

Fig. 6. Specificity.

Specificity metric evaluation and its comparison is illustrated in above shown Fig. 6. In that, various techniques are represented in x-axis and specificity value is expressed in y-axis. Available methods are such as RNN with C3D and CNN algorithms provide lower specificity whereas DCNNDHO algorithm provides higher specificity for the UCF and kinetics datasets. Thus the result concludes that the DCNNDHO increase the video classification accuracy for the both datasets.

Recall and Precision values weighted average defines F1 score. Both false negatives and false positives are considered in this score computation.

$$\text{F - measure} = 2 * \frac{(\text{sensitivity} * Precision)}{(\text{sensitivity} + Precision)} \tag{5}$$

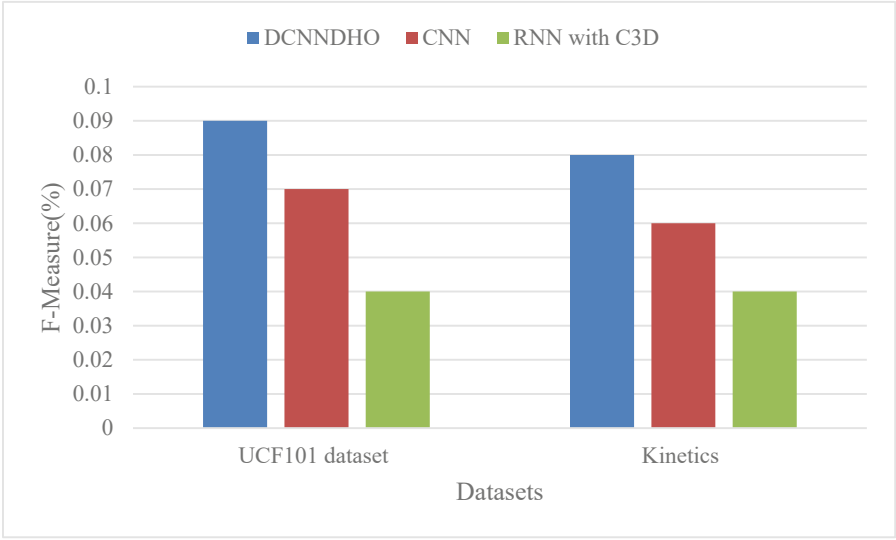

Fig. 7. F-measure.

F-measure metric evaluation and its comparison is illustrated in above shown Fig. 7. In that, various techniques are represented in x-axis and F-measure is expressed in y-axis. Available methods are such as RNN with C3D and CNN algorithms provides low F-measure whereas DCNNDHO algorithm provides high F-measure for UCF and kinetics datasets. Thus the result concludes that the DCNNDHO increase the video classification accuracy for the both datasets.

3.5 False Discovery Rate

Ratio between false positive results count to total positive test results count defines false discovery rate.

$$\text{False discovery rate} = V/V + S \tag{6}$$

where, false discoveries count is represented as V and true discoveries count is represented as S.

False discovery rate metric evaluation and its comparison is illustrated in above shown Fig. 8. In that, various techniques are represented in x-axis and false discovery rate is expressed in y-axis. Available methods are such as RNN with C3D and CNN algorithms provide higher false discovery rate whereas DCNNDHO algorithm provides lower false discovery rate for the UCF and kinetics datasets. Thus the result concludes that the DCNNDHO increase the video classification accuracy for the both datasets.

3.6 False Alarm Rate

Ratio between false alarms count to total alarms or warning count defines thus false alarm rate.

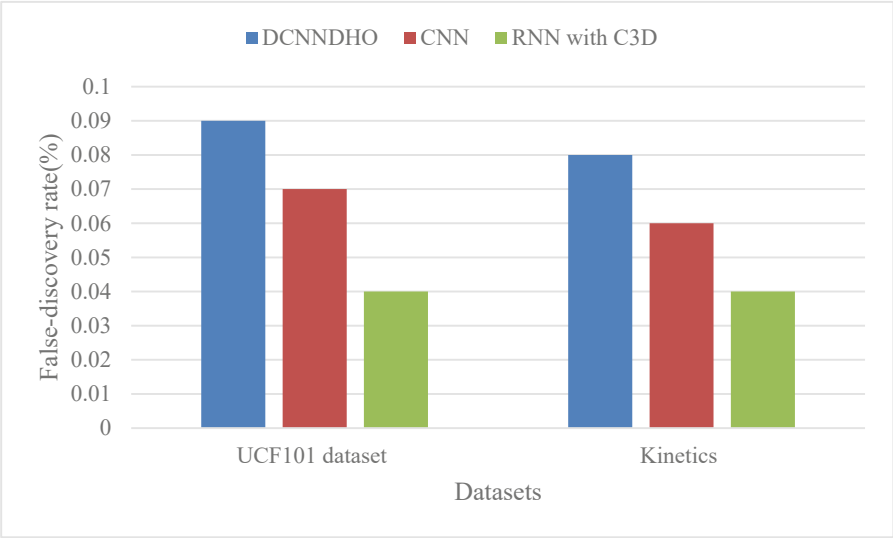

Fig. 8. False discovery rate.

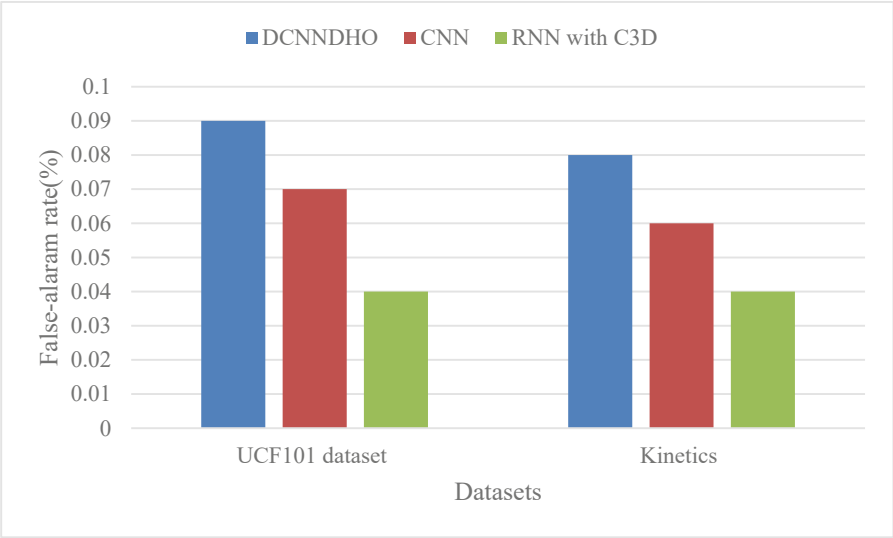

Fig. 9. False alarm rate.

False alarm rate metric evaluation and its comparison is illustrated in above shown Fig. 9. In that, various techniques are represented in x-axis and false alarm rate is expressed in y-axis. Available methods are such as RNN with C3D and CNN algorithms

provide higher false alarm rate whereas DCNNDHO algorithm provides lower false discovery rate for the UCF101 dataset, and Kinetics dataset. Thus the result concludes that the DCNNDHO increase the video classification accuracy for the both datasets.

4 Conclusion

This study examined video content analysis using segmentation, recognition, and classification approaches. Several feature extraction, segmentation, and classification algorithms are analyzed on video content to increase performance. Feature extraction is a crucial stage in extracting essential information from video material. Texture, motion, and color information are extracted for classifier training. Then the video segmentation is optimized using DGA and deep learning methods. Thus, picking characteristics improves classification accuracy. CNN, MLP, SVM, naive Bayes, and KNN are used to classify the dataset. Each algorithm's flaws and benefits are explored. This poll provides a helpful summary of available solutions, with their drawbacks and benefits. However, existing algorithms struggle to identify key aspects and their strengths and limitations in video content classification. The DCNNDHO algorithm is used to increase video classification accuracy. This reduces categorization errors. This improved classifier finds biases and weights between hidden and input neurons. However, deep learning algorithms are not efficient for video content classification in this survey. Hybrid optimization techniques can also be presented to improve video content detection and classification.

References

1. Gaunt, K.D.: YouTube, twerking & you: context collapse and the handheld co-presence of black girls and Miley Cyrus. J. Popular Music Stud. **27**(3), 244–273 (2015). ISBN 9781315689593
2. Loukas, C.: Video content analysis of surgical procedures. Surg. Endosc. **32**(2), 553–568 (2018). https://doi.org/10.1007/s00464-017-5878-1
3. Bai, L., et al.: Video semantic content analysis based on ontology. In: International Machine Vision and Image Processing Conference (IMVIP 2007). IEEE (2007). https://doi.org/10.1109/IMVIP.2007.13
4. Perazzi, F., et al.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016). https://doi.org/10.1109/CVPR.2016.85
5. Zhang, N., et al.: A generic approach for systematic analysis of sports videos. ACM Trans. Intell. Syst. Technol. **3**(3) (2012). Article 46
6. Cricri, F., et al.: Sport type classification of mobile videos. IEEE Trans. Multimedia **16**(4), 917–932 (2014)
7. Chen, Y., et al.: Accelerating video feature extractions in CBVIR on multi-core systems. Intel Technol. J. **11**(4) (2007). https://doi.org/10.1535/itj.1104.08. ISSN 1535-864X
8. Shih, H.-C., Huang, C.-L.: Content extraction and interpretation of superimposed captions for broadcasted sports videos. IEEE Trans. Broadcast. **54**(3), 333–346 (2008). https://doi.org/10.1109/TBC.2008.2001143
9. Pang, Y., Yan, H., Yuan, Y., Wang, K.: Robust CoHOG feature extraction in human-centered image/video management system. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **42**(2), 458-468 (2012)

10. Cambria, E., Poria, S., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
11. Xu, C., et al.: Visual sentiment prediction with deep convolutional neural networks (2014)
12. You, Q., et al.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)
13. Tran, D., et al.: Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
14. Poria, S., et al.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th International Conference on Data Mining (ICDM) (2016)
15. Yong, S.-P., Deng, J.D., Purvis, M.K.: Wildlife video key-frame extraction based on novelty detection in semantic context. Multimedia Tools Appl. **62**(2), 359–376 (2013)
16. Kantor, V., Laptev, I.: Efficient feature extraction, encoding and classification for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2593- 2600 (2014). https://doi.org/10.1109/CVPR.2014.332
17. Zhang, W., Duan, P., Lu, Q., Liu, X.: A realtime framework for video object detection with storm. In: Ubiquitous Intelligence and Computing, 2014 IEEE 11th International Conference on and Autonomic and Trusted Computing, IEEE 14th International Conference on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom), pp. 732–737 (2014). https://doi.org/10.1109/UIC-ATC-ScalCom.2014.115
18. Chen, C.-M., Chen, L.-H.: A novel approach for semantic event extraction from sports webcast text. Multimedia Tools Appl. **71**(3), 1937–1952 (2012). https://doi.org/10.1007/s11042-012-1323-6
19. Chivadshetti, P., Sadafale, K., Thakare, K.: Content based video retrieval using integrated feature extraction and personalization of results. In: 2015 International Conference on Information Processing (ICIP). IEEE (2015). https://doi.org/10.1109/INFOP.2015.7489372
20. Zhang, Z., et al.: Efficient video frame insertion and deletion detection based on inconsistency of correlations between local binary pattern coded frames. Secur. Commun. Netw. **8**(2), 311–320 (2015). https://doi.org/10.1002/sec.981
21. Zhao, L., Wang, Z., Zhang, G.: Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and Gabor multiorientation fusion histogram. Math. Probl. Eng. **2017**, 12. Article ID 7206041. https://doi.org/10.1155/2017/7206041
22. Abdulhussain, Sadiq H., et al. "A fast feature extraction algorithm for image and video processing." 2019 international joint conference on neural networks (IJCNN). IEEE, 2019. DOI: https://doi.org/10.1109/IJCNN.2019.8851750
23. Kim, E.Y., Park, S.H.: Automatic video segmentation using genetic algorithms. Recogn. Lett. **27**(11), 1252–1265 (2006). https://doi.org/10.1016/j.patrec.2005.07.023
24. Zhu, S., Liu, Y.: Video scene segmentation and semantic representation using a novel scheme. Multimed. Tools Appl. **42**, 183–205 (2009). https://doi.org/10.1007/s11042-008-0233-0
25. Budvytis, I., Badrinarayanan, V., Cipolla, R.: Semi-supervised video segmentation using tree-structured graphical models. In: CVPR 2011. IEEE (2011). https://doi.org/10.1109/CVPR.2011.5995600
26. Yu, C.-P., et al.: Efficient video segmentation using parametric graph partitioning. In: Proceedings of the IEEE International Conference on Computer Vision (2015). https://doi.org/10.1109/ICCV.2015.361
27. Piramanayagam, S., Saber, E., Cahill, N.D.: Gradient-driven unsupervised video segmentation using deep learning techniques. J. Electron Imaging **29**(1), 013019 (2020). https://doi.org/10.1117/1.JEI.29.1.013019
28. Huang, C., Tianjun, F., Chen, H.: Text-based video content classification for online video-sharing sites. J. Am. Soc. Inform. Sci. Technol. **61**(5), 891–906 (2010)

29. Fan, Y., et al.: Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: ICMI 2016: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 445–45010 (2016). https://doi.org/10.1145/2993148.2997632
30. Ullah, A., et al.: Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access **6**, 1155–1166 (2017). https://doi.org/10.1109/ACCESS.2017.2778011
31. Marsden, M., et al.: ResnetCrowd: a residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE (2017). https://doi.org/10.48550/arXiv.1705.10698
32. Aljarrah, I., Mohammad, D.: Video content analysis using convolutional neural networks. In: 2018 9th International Conference on Information and Communication Systems (ICICS). IEEE (2018). https://doi.org/10.1109/IACS.2018.8355453
33. Hossain, M.S., Muhammad, G.: Emotion recognition using deep learning approach from audio-visual emotional big data. Inf. Fusion **49**, 69–78 (2019). https://doi.org/10.1016/j.inffus.2018.09.008
34. Shri, S.J., Jothilakshmi, S.J.C.C.: Crowd video event classification using convolutional neural network. Comput. Commun. **147**, 35–39 (2019). https://doi.org/10.1016/j.comcom.2019.07.027
35. Minhas, R.A., et al.: Shot classification of field sports videos using AlexNet convolutional neural network. Appl. Sci. **9**(3), 483 (2019). https://doi.org/10.3390/app9030483
36. Khan, M.N., Ahmed, M.M.: Snow detection using in-vehicle video camera with texture-based image features utilizing K-nearest neighbor, support vector machine, and random forest. Transp. Res. Rec. **2673**(8), 221–232 (2019). https://doi.org/10.1177/0361198119842105
37. Shambharkar, P.G., Doja, M.N.: Movie trailer classification using deer hunting optimization based deep convolutional neural network in video sequences. Multimedia Tools Appl. **79**(29–30), 21197–21222 (2020). https://doi.org/10.1007/s11042-020-08922-6
38. Sreekanth, N., SasiKiran, J., Obulesu, A., Mallikarjuna Reddy, A.: key frame extraction for content based lecture video retrieval and video summarisation framework. European J. Mol. Clin. Med. **7**(11), 496–507 (2020). ISSN 2515-8260
39. Jai Shankar, B., Murugan, K., Obulesu, A., Finney Daniel Shadrach, S., Anitha, R.: MRI image segmentation using bat optimization algorithm with fuzzy C means (BOA-FCM) clustering. J. Med. Imaging Health Inform. **11**(3), 661–666 (2021)
40. Obulesh, A., et al.: Central nervous system tumour classification using residual neural network, Purakala. UGC Care J. **31**(21) (2020). ISSN 0971-2143
41. Obulesh, A., et al.: Traffic-sign classification using machine learning concepts, Tathapi. UGC Care Listed J. **19**(8) (2020). ISSN 2320-0693
42. Thimmaraju, R., Obulesh, A., Reddy, M.S.: Quantum computation and simulation a distinguished demonstration using the BruteForce algorithm. In: 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1–6. IEEE (2021). https://doi.org/10.1109/INOCON50539.2020.9298345
43. An, G., Zheng, Z., Wu, D., Zhou, W.: Deep spectral feature pyramid in the frequency domain for long-term action recognition. J. Vis. Commun. Image Represent. **64**, 102650 (2019)
44. Xiao, J., Cui, X., Li, F.: Human action recognition based on convolutional neural network and spatial pyramid representation. J. Vis. Commun. Image Represent. **71**, 102722 (2020)
45. Tiger, M., Heintz, F.: Incremental reasoning in probabilistic signal temporal logic. Int. J. Approximate Reasoning **119**, 325–352 (2020)