International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Deep Learning for Dense Interpretation of Video: Survey of Various Approach, Challenges, Datasets and Metrics

Kiran P Kamble1

Department of Technology Shivaji University,

Kolhapur Maharashtra, India -416004

kirankamble5065@gmail.com

Vijay R. Ghorpade2

Department of Computer Science & Engineering,

BVCOEK, Shivaji University, Kolhapur

vijayghorpade@rediffmail.com

Abstract— Video interpretation has garnered considerable attention in computer vision and natural language processing fields due to the rapid expansion of video data and the increasing demand for various applications such as intelligent video search, automated video subtitling, and assistance for visually impaired individuals. However, video interpretation presents greater challenges due to the inclusion of both temporal and spatial information within the video. While deep learning models for images, text, and audio have made significant progress, efforts have recently been focused on developing deep networks for video interpretation. A thorough evaluation of current research is necessary to provide insights for future endeavors, considering the myriad techniques, datasets, features, and evaluation criteria available in the video domain. This study offers a survey of recent advancements in deep learning for dense video interpretation, addressing various datasets and the challenges they present, as well as key features in video interpretation. Additionally, it provides a comprehensive overview of the latest deep learning models in video interpretation, which have been instrumental in activity identification and video description or captioning. The paper compares the performance of several deep learning models in this field based on specific metrics. Finally, the study summarizes future trends and directions in video interpretation.

Keywords- Video Interpretation, Deep Learning, Video Datasets, Video Features

Introduction

The volume of digital multimedia data (text, image, audio, and video content) that is created and shared is rising exponentially. With the growth of sensors and mobile devices, video has become a widespread communication medium or transmission module among Internet users. This tremendous rise in video is due to advancements in transmission technology, capturing devices, and display methods. Every minute, thousands of hours of video are posted to YouTube and Facebook, which must be swiftly comprehended. This problem may be solved by using automatic caption generation to describe images and videos. Recently, video comprehension has become a major focus of study. Video description in plain language is easy for people, but very difficult for computers. As a result, in order for video captions to be relevant they must be able to grasp what is going on in a video in terms of objects, interactions, the spatio-temporal sequence of events, and other such minutia. Detection and generation of description from images and videos is one of the major tasks in image processing and computer vision. Applications such as video comprehension, human computer

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

interaction, automated video subtitling, and aiding the visually challenged folks rely on video captioning or description [2]. Work in this area has grown at a rapid pace in the last several years.

Natural language processing (NLP) and understanding of visual contents have never been linked before. Visual content and language learning have been seen as a tough job and a vital step toward machine intelligence with various applications in everyday settings such as image/video retrieval, video comprehension, blind navigation, and automated video subtitling [1]. Video captioning is an important step toward artificial intelligence. In addition, it is capable of bridging the gap between visual and verbal communication and aiding those with visual impairments in understanding video information. To put it another way, it opens the door to a wider range of possibilities, such as the ability to transform a collection of relevant films into a report page. Because of its importance in applications that function in a real-time environment based on activity detection, activity recognition is one of the most promising tasks.

Video Description, particularly, goes beyond video captioning to offer a more complete analysis of the video's visual elements. Video captioning, on the other hand, is more challenging than picture captioning due to the wide range of objects, scenes, actions, qualities, and prominent elements. Despite the complexity of video captioning, a few efforts have been made, mostly motivated by current deep learning developments. Video interpretation includes various subclasses namely video description, video captioning, activity recognition, video summarization, dense image description. In this paper, we presented a comprehensive survey on recent deep learning models in the above-mentioned subclasses, datasets used, feature extraction techniques, evaluation metrics in video interpretation. Figure 1 illustrates the path we followed for the detailed survey, while Figure 2 provides an example of automatic video content description.

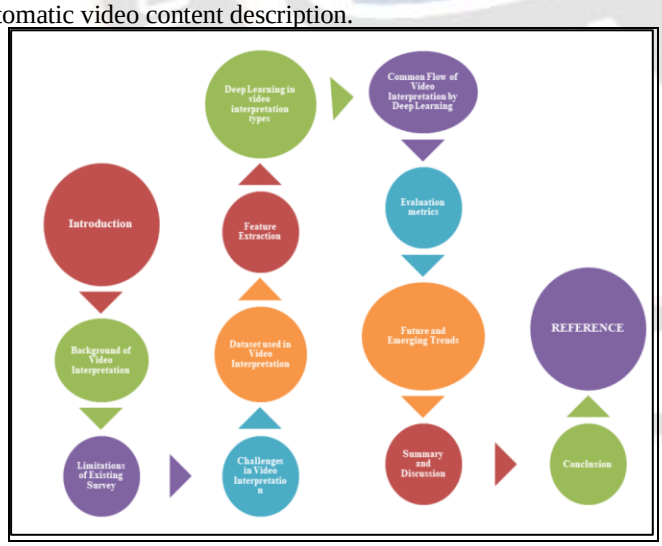

Figure 1: Path adopted for detail survey.

I. Background of Video Interpretation

Video accessibility is essential for the education, employment, and helping those who are visually impaired. About 285 million people worldwide are visually impaired and 39 million individuals are blind, according to the World Health Organization (WHO). Many videos on the Internet are not accessible to those who are blind or visually impaired, despite international rules and regulations. Professional video explanations are expensive and time demanding to produce. However, the quality of the volunteer-created video descriptions might vary widely, making them an unattractive option for those who are new to the profession of video description [3]. Consequently, there is a need to automate video text production in computer vision.

Recent developments in Computer Vision have made automatic image description creation an intriguing but challenging issue. The classical era of visual description research used conventional approaches such as video description and natural

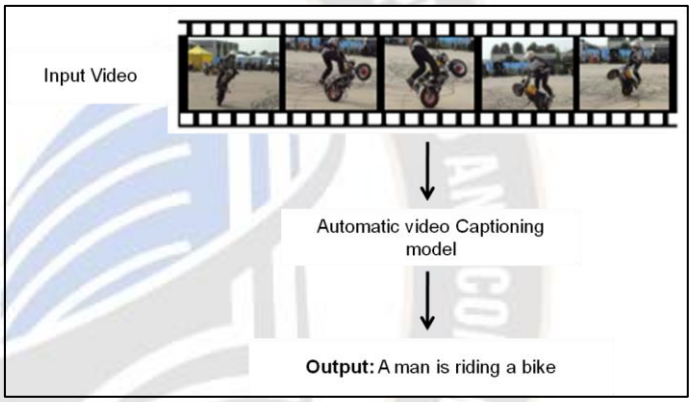

Figure 2: An example of automatic video content description

language processing (NLP) to first recognize items in films and then fit them to standard sentence templates. We are now at a point in the research process where we are confident that we can solve the open domain automated video description issue with deep learning. In several computer vision issues, such as object identification, object detection, and activity recognition, major developments in Convolutional Neural Network (CNN) models have made this feasible. Video description generation is an even more difficult problem that might have several applications for multimedia applications or for blind people or human-robot interaction.

A. Traditional Video Interpretation methods

Language templates have been used in most approaches to handle video as a flat data sequence, neglecting its inherent multimodality nature. Starting with template-based techniques, researchers in video captioning began by combining SVOs (Subject, Verb, and Object) together using a phrase template that had been extracted from the video [4]. SVO-triplets are the term used to describe this kind of operation.

SVO triplets (subject-verb-object) are extracted from training sentences and matched to visual data during the test phase of visual content description in the early stages of the process. On

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

the other hand, the lack of variation in produced sentences as well as extremely dependent on the syntactical structures of templates are both drawbacks of this bottom-up method (Bin et al. 2018) [5]. In addition, a lack of grammatical resources might lead to an incorrect description being created. For example, a guy is performing on a stage rather than a man performing on stage. SVO triplets' representations, on the other hand, are unable to convey video's temporal information (such as successive actions). These approaches were superseded by deep learning when huge datasets showed that they couldn't handle open domain video diversity.

There are several advantages to using a top-down approach to learning video representation encoder and phrase decoder concurrently, which is inspired by the neural translation machine's success in achieving a high level of translation accuracy. A top-down workflow typically uses convolutional neural networks (CNNs) to retrieve static image features frame-by-frame, and then combines the features of all frames into one global representation for the video using various kinds of operations. In order to transform a visual representation into a phrase word-by-word, RNNs are then used.

III. Limitations of Existing Survey

Islam et al. 2021 [8] offered an assessment on state-of-the-art methodologies, emphasizing deep learning models, analyzing benchmark datasets in multiple parameters, and grading the advantages and drawbacks of the different evaluation metrics based on the prior research in the video captioning area. This survey provides a detailed description of limited datasets. Aafaq et al. 2019 [9] presented a review on state-of-the-art DL techniques, benchmark datasets, and assessment criteria in video description. Rafiq et al. 2021 [10] offered a survey that focuses solely on the benchmark datasets, and assessment metrics established and implemented for video description jobs and their capabilities and limits. This survey does not discuss about the datasets. In the existing surveys, challenges posted by corporations on video interpretation like activitynet challenges are not described. In this paper, we presented a comprehensive survey on deep learning models in video interpretation (video captioning and description, video summarization, activity recognition, dense image description), challenges posted by corporations on video interpretation in recent years, various datasets, feature extraction techniques, and evaluation metrics. We presented the comparison of several datasets including ActivityNet Caption, HMDB51, Hollywood 2, TREC, TACOS, KTH, MPII-MD, M-VAD, MSRVTT, MSVD, YouCook, TACOS Multilevel, Sports 1M, THUMOS, UCF50, and UCF 101 in this survey. Discussion on different metrics, for evaluating video interpretation models, such as BLEU, METEOR, ROUGE, CIDEr, rank based measures, WMD, Semantic Textual Similarity, Direct Assessment, PR15, and nAUDC is provided here.

IV. Challenges in Video Interpretation

Units Our understanding of this link between visual content and sentence semantics has not previously been examined in video description. Ideally this relationship should be modelled and integrated in natural language. In video description, the technique of "temporal attention" has been frequently utilised to selectively concentrate on key frames. As a result, many current systems relying on temporal attention suffer from issues such as incorrect recognition or a lack of information. The work of machine learning algorithms in video content understanding still has many problems and faces many different challenges. Different grand challenges on video understanding are posted every year by corporations.

A. ActivityNet

ActivityNet is a new large-scale video standard for the study of human activities. Singh and Vishwakarma 2019 used a flexible structure that allows consistent acquisition, crowdsourced annotation and segmentation of online videos. This results in the ActivityNet activity dataset, which is large in terms of the number of categories and samples per category, rich in taxonomy, and simple to use. Activities are organized by social interaction and location in ActivityNet, which is one of the most essential features of the system. There are at least four layers of activity hierarchy provided by this tool.

ActivityNet challenges 2021 holds 12 diverse challenges (described in table 1), with the goal of pushing the boundaries of semantic visual interpretation of movies and linking visual material with human captions as they go forward. Time-based evidence in the form of class labels, captions, and object entities is what these assignments concentrate on.

TABLE 1: TASKS IN ACTIVITYNET CHALLENGES 2021

ActivityNet Challenges 2021	**Description**
Kinetics 700 Challenge	It focused on both supervised and self-supervised video classification.
TinyActions Challenge	The focus is on recognizing tiny actions in videos.
ActivityNet Temporal Action Localization Challenge 2021	This project aims to see how well algorithms can find events in untrimmed video sequences.
HACS Temporal Action Localization Challenge	The purpose of this challenge is to temporally locate activities in untrimmed videos in supervised and weakly-supervised strategies.
SoccerNet Challenge	The SoccerNet-V2 dataset, which covers over 500 games from three seasons of the six main European football leagues, is used in this challenge. Given a professional

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

	soccer broadcast, it seeks to motivate players to recognize the precise timestamps in the video at which specific acts occur, and replayed events.
AVA-Kinetics & Active Speakers	This challenge aims to solve two key problems in spatiotemporal video comprehension such as locating action extents in space and time, and detecting active speakers in video sequences in a dense manner.
ActEV SDL Unknown Facility (UF)	It supports the use of an ActEV Command Line Interface (CLI) and submission to create algorithms to identify and temporally locate incidents of Known activities.
ActivityNet Event Dense-Captioning	Video events must be identified as well as described to complete this job.
ActivityNet Entities Object Localization	The goal of this exercise is to determine how accurate a description (generated or ground-truth) is in relation to the video it describes.
Video Semantic Role Labeling	VidSRL has three sub-tasks including prediction of a verb sense that describes the most important action, prediction of given verb's semantic roles, and prediction of event relations.
MMAct Challenge	This challenge concentrates on cross-modal understanding of video actions strategies to overcome the limits given by the modality disparity between the training and testing phases by using both sensor- and vision-based modalities in ways that addressed the drawbacks imposed by visual-only approaches.
HOMAGE (Home Action Genome)	It focuses on recognizing compositional activity in the house, but it also includes numerous viewpoints and more sensor modalities.

B. DCEV-ActivityNet

It is more challenging to extract features because of the similarity across activities. The qualities of several activities may be similar (e.g., walking and running). In order to depict activities in a unique way, distinct traits are tough to come by. Multi-activity movies are available in a variety of datasets. With a single phrase caption, it is impossible to convey several scenes or activities in a video. Understanding that various modalities and the constituents inside each modality have varying effects on the creation of sentences. It's difficult to come up with a large number of associated phrases at once. Using many phrases relevant to the video's whole context, dense video captioning was able to solve this difficulty (Dave and Padmavathi 2022). Dense-Caption Events in Video (DCEV) is proposed to generate event proposals and context-based caption generator to generate captions. ActivityNet challenge 2018's extensive video captioning is often broken down into two stages: In order to identify probable events in the video, two methods are used: (1) event proposal generation and (2) event caption creation. Detecting and characterizing the occurrences in a video are both part of the dense-captioning job studied in this challenge. Use the ActivityNet Captions dataset, a new standard for densely captioned events, to compete in this competition.

C. Large Scale Movie Description Challenge (LSMDC)

LSMDC was presented by Rohrbach et al. 2017 [6] and provides a parallel corpus of 128,118 phrases matched to video clips from 200 movies. Film clips will be described automatically in the challenge. By integrating the “M-VAD and MPII-MD” datasets, they created LSMDC In order to avoid having the same movie show up in the combined dataset, they first determined the overlap between the two. For validation and testing, they also omitted script-based movie alignments. “LSMDC 2016” comprises 101,046 training clips, 7408 validation clips, and 128K total clips after manual alignment of training and validation sets.

As of 2019, LSMDC's newest challenge tracks seek to generate multi-sentence movie descriptions in a more realistic and practical context. On the basis of a group of five snippets, movie summaries are judged rather than individual clips. The importance of distinguishing "who is who" while narrating a series of events cannot be overstated. As a result, the emphasis of the challenge will be on character identification rather than generic "SOMEONE". Table 2 shows the summary of LSMDC challenges and winners of these challenges from 2015 to 2019.

TABLE 2: LIST OF LSMDC CHALLENGES AND WINNER

LSMDC challenge	**Task**	**Winner**
LSMDC 2015		“Video Captioning with Recurrent Networks Based on Frame- and Video-

		Level Features and Visual Content Classification" by "Rakshith Shetty and Jorma Laaksonen"
LSMDC 2016	Movie Description	"Video Description by Combining Strong Representation and a Simple Nearest Neighbor Approach" by "Gil Levi, Dotan Kaufman, Lior Wolf, and Tal Hassner".
	Movie Annotation and Retrieval	"Video Captioning and Retrieval Models with Semantic Attention" by "YoungJae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim".
	Movie fill-in-the-blank	"Video Captioning and Retrieval Models with Semantic Attention" by "YoungJae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim".
LSMDC 2017	Movie Description	"MTLE: A Multitask Learning Encoder of Visual Feature Representations for Video and Movie Description" by "Oliver Nina, Scott Clouse, and Alper Yilmaz".
	Movie Annotation and Retrieval	"Multi Sequence to One : Joint Sequence Fusion Model for Video Question-Answering and Retrieval" by "YoungJae Yu, Jongseok Kim, Gunhee Kim"[99]
	Movie fill-in-the-blank	"Multi Sequence to One : Joint Sequence Fusion Model for Video Question-Answering and Retrieval" by "YoungJae Yu, Jongseok Kim, Gunhee Kim"
LSMDC 2019	Multi-Sentence Description	"Auxiliary Loss assisted Multi-sentence generation" by "Youngjae Yu, Jiwan Chung, Jongseok Kim, Heeseung Yun, and Gunhee Kim"
	Fill-in the Characters	"Story-Character Matching Network" by "Youngjae Yu, Jiwan Chung, Jongseok Kim, Heeseung Yun, and Gunhee Kim"
	Multi-Sentence Description with Characters	Combination of above two approaches

D. Microsoft multimedia challenge MSR-VTT

For decades, video recognition has been a major problem for computer vision researchers. In the past, researchers have mostly concentrated on detecting films using a preset, but very restricted, collection of individual words. Video to text (VTT) translation is the goal of Microsoft Research-VTT's (MSR-VTT) great challenge. The objective is to automatically construct a comprehensive and natural phrase to explain video information, preferably containing its most interesting characteristics, from a given video clip that has been entered. The context is determined by semantics rather than temporal attention. Semantic and temporal attention have never been used together in video captioning before. Fused GRU with Semantic-Temporal Attention (STA-FG) is a pipeline developed by Gao et al. 2020 that uses the MSR-VTT dataset to generate semantic-temporal attention for video captioning while explicitly including high-level visual notions.

E. National Institute of Standards and Technology (NIST) sponsored challenge TRECVID

With the NIST-TREC Video Retrieval Evaluation (TRECVID) project, researchers want to create bigger and larger multicamera datasets that may be utilized to plan future Activities in Extended Video (ActEV) tasks. Computer vision researchers are encouraged to create new algorithms that can better identify human activity in multi-camera systems that span a vast area in the NIST-TREC Video Retrieval Evaluation ActEV Challenge. Table 3 illustrates the summary of TRECVID challenges of recent years.

TABLE 3: SUMMARY OF RECENT TRECVID CHALLENGES

NIST-TRECVID Challenge	**Tasks**	**Definition**
TRECVID 2019	AVS (Ad-hoc Video Search)	Introduction of Vimeo Creative Commons Collection (V3C) new dataset
	ActEV	MI dataset
	INS (Instance Search)	BBC EastEnders soap opera episodes
	VTT (Video to Text)	Vine video data
TRECVID 2020	AVS	V3C dataset
	INS	BBC EastEnders soap opera episodes]
	VTT	V3C dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

	ActEV	VIRAT dataset
	VSUM (Video Summarization)	BBC EastEnders soap opera episodes
	DSDI (Disaster Scene Description and Indexing)	Low Altitude Disaster Imagery
TRECVID 2021	Tasks similar to TRECVID 2020	

V. DATASET USED IN VIDEO INTREPRETATION

The rapid growth of this study field is largely due to the availability of labelled datasets for video interpretation. We compared and contrasted distinct sets of data in our study.

A. ActivityNet Caption

There are 20k untrimmed movies in the ActivityNet-Caption dataset, each of which has an average of 3.6 event clips with captions, separated into training, validation and testing subsets by a 2:1:1 ratio (Yu and Han 2021). Videos from 200 classes are included in ActivityNet-Captions, which is based on the ActivityNet platform. More than three annotated events with human-written captions can be found in every film, which is around 120 seconds long on average.

B. HMDB 51

For video action detection, the HMDB51 dataset comprises real-world movies and online videos that are more complex in visual content and hence provide a greater challenge. Each of the 51 action classes in HMDB51 has at least 100 video clips, and there are 6,766 clips in total (Xu et al. 2018).

C. Hollywood 2

The Hollywood 2 Dataset comprises 3,669 samples from 69 Hollywood films, including 12 action categories and 10 scenes (Ding et al. 2019). Due of their preference for particular types of material (movies and sports), Hollywood-2 has the highest center bias (Droste et al. 2020). Analysis or detection of actors in Hollywood-2 video samples is difficult because to the wide range of emotions, postures, clothing, camera movements, lighting changes, occlusions, and backdrops, all of which are comparable to those seen in actual settings.

D. TREC

A selection of movies from the 2007/8 TREC video assessments was made by Dilawari et al 2018. In all, there are seven different types of data: activities, close-ups (news stories), meetings (groups), traffic, and scenes. In all, there are 140 videos in total, 20 in each category. Each movie is between 10 and 30 seconds in duration. TRECVID doesn't provide any written annotations for the video segments.

E. TACoS

Textually Annotated Cooking Scenes (TACoS) dataset is used for grounding and dense video captioning jobs in video grounding and dense video captioning. There are 127 videos for a total of 4,79 minutes in duration. The TACoS dataset comprises 18818 video-query pairings for the video grounding task. TACoS contain more video segments with queries per video than ActivityNet Captions dataset. An average of 148 queries is run through each video. This makes TACoS dataset very tough, since searches cover just few seconds or even many frames, making it extremely time-consuming to analyses (Zeng et al. 2020).

F. KTH

The KTH is the most often used publicly available dataset on human behaviour. It has a resolution of 160 by 120 pixels and six different sorts of video activities (Pan and Li 2020) (Jaouedi et al. 2020). The basic surroundings, lack of camera movement, limited number of motions, and the presence of just one person in every movie with a single movement are all clear drawbacks of the KTH datasets. These datasets depict scenarios that are quite different from those that exist in the actual world.

G. MPII-MD

Movie datasets like the "MAX PLANK INSTITUTE FOR INFORMATICS - MOVIE DESCRIPTION (MPII-MD)" are often used. The average duration of each movie is 3.9 seconds, or about one line, and the transcripts for 94 Hollywood films are included (Sun et al. 2021). Almost 73.6 hours of video content and 653,467 words make up the dataset. There are 56,861 training videos, 4,930 validation videos, and 6,584 testing videos.

H. M-VAD

There are 48,986 video clips in the Montreal Video Annotation Dataset (M-VAD) that were taken from 92 movies. An average of 6.2 seconds is spent on each clip, and there are 55,904 phrases in total (Saleem et al. 2019). To put it another way, there are 84.6 hours' worth of video clips in the collection. Video clips with more than one phrase make up the majority of the 55,904 sentences available. Data for training and validation is 38,949 videos and testing data is 4,888 videos.

I. MSRVTT

In all, there are 10,000 YouTube videos and 200k descriptors in this dataset. The videos are 14 seconds long on average. Multiple humans provide annotations to each video (Tu et al. 2021).

J. MSVD

Among the themes covered by the MSVD dataset for video captioning include sports, wildlife, and music. It includes 1,970 YouTube videos. Approximately 8,000 English descriptions are available, with an average of 40 explanations for each movie (Jin et al. 2019).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

K. YouCook

One hundred and eighty-eight YouCook recipes from a wide range of different YouTubers are included in the dataset (Shin et al. 2022). In majority of the videos, the kitchen/scenery is different. Compared to the MP-II Cooking dataset, which was filmed with a fixed camera perspective in the same kitchen and with the same backdrop, this dataset poses a more difficult visual difficulty. Grilling, baking, and so on are among the six culinary methods included in the dataset. The training set for machine learning includes 49 videos, whereas the test set includes 39 videos.

L. TACoS multilevel

There are many levels of depth in the description provided by TACoS-MultiLevel (Bhowmik et al. 2021). For each of the three types of descriptions, workers were instructed to use no more than 15 words and no less than three to five phrases, and no more than one sentence for each kind of description. Each video's description is represented by around 20 triples in the dataset.

M. Sports 1M

One million YouTube videos belonging to 487 sports classifications are included in Sports1M, another large-scale video collection (Zhang et al. 2019). Classes including boxing, soccer, and volleyball may be found here. There is video annotation for the complete untrimmed movie, but the temporal bounds of the acts are unknown (Poorgholi et al. 2021).

N. THUMOS

THUMOS 14 dataset contains videos across 20 categorizes of sports classes (Gao et al. 2020) [7]. Long uncut films may be found on THUMOS, however the most of them (85%) only have one action class (Yeung et al. 2018) [30].

O. UCF 50

The UCF-50 dataset comprises footage of 50 acts in unrestricted settings. There are 6700 videos in total, with an average of 100-150 films each category. The UCF-11 dataset is a subset of this one (Roy et al. 2021) [31].

P. UCF101

With at least 100 video clips for each of the 101 action courses, UCF101 has a total of 101 classes. It's broken down into 25 categories based on the kind of performance. This dataset contains 13,320 clipped video segments (Zuo et al. 2019) [32].

TABLE 4: DATASET USED IN VIDEO INTERPRETATION

Dataset	**Domain**	**# classes**	**#videos**	**#avg len**	**#clips**	**#sent**	**#words**	**#vocab**	**#len (hrs)**
MSVD	open	219	1971	11s	1,971	71,028	607,339	13,010	5.3
MPII Cooking	cooking	66	45	601 s	-	5,608	-	-	7.1
YouCook	cooking	7	87	-	-	2,687	41,458	3,712	2.4
TACoS	cooking	25	128	361 s	7,205	17,227	146,771	28,292	15.9
TACos-MLevel	cooking	2	186	361 s	15,105	52,593	2K	-	28.1
MPII-MD	movie	2	95	3.8 s	67,337	68,375	653,467	24,549	73.6
M-VAD	movie	1	91	6.3 s	49,986	55,904	519,933	17,609	84.6
MSR-VTT	open	21	7,182	21s	11K	200K	1,856,523	29,316	41.2
YouCook II	cooking	88	1K	317 s	16.5 K	16.5 K	-	2,602	176.2
ActivityNet Captions	Human activity	200	20K	120 s	-	-	13.65	-	648

VI. FEATURE EXTRACTION

The visual information in real-world online videos is typically supplemented with clues that may be used to generate natural language descriptions. Video data's rich temporal information, which may be used for activity analysis in addition to geographical distribution research, is a unique advantage (Mou and Zhu 2016) [33]. Visual words are quantized local descriptors used to describe images and videos. The visual and linguistic modalities are not effectively represented by these hand-crafted elements; therefore, they cannot be directly compared. A common latent subspace is therefore discovered where the two modalities may be more accurately represented and an estimate of their similarity can be derived (Dong et al. 2018). Two major categories of video features are Single Stream (either spatial or temporal) and two stream features. The spatial stream recognizes activity from static video frames, but the temporal stream is taught to recognize action from intense optical flow motion. For video interpretation, two stream characteristics incorporate both spatial and temporal information. As shown in table 5, video interpretation using two stream features is more accurate than single stream features. Table 6 illustrates the merits and demerits of video interpretation features.

TABLE 5: ASSESSMENT OF MAIN FEATURE CATEGORIES IN VIDEO CAPTIONING (SUN ET AL. 2021) [108]

Features		**Accuracy (%)**
Single Stream Features	Spatial Features	73
	Temporal Features	83

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Two Stream Features (Spatial + Temporal features)	87

A. ResNet-50

Choi et al. 2021 investigated whether two models, "Vanilla-RNN with ResNet50" and "Bi-directional RNN with ResNet50", can prepare sequential data in a stable manner while considering RNN consequences. They analyzed the effects of feature extraction of these two models. For action recognition, Suresh and Visumathi 2020 proposed a novel deep neural network architecture that uses transfer learning. Convolutional neural networks (CNNs) and a long-term model were used to build the model (LSTM). Extracting feature vectors from Inception ResNet v2 is used to train the model. For video captioning, Hammad et al. 2020 used ResNet 50 to extract scene recognition characteristics. In order to extract visual elements needed for caption development, Lee and Kim 2018 used ResNet. Figure 2 shows the results of feature extraction using ResNet 50 as the training network. ResNet-based features outperform GoogleNet and VggNet-based features in terms of performance (Song et al. 2018). Figure 3: Feature Extraction using ResNet-50

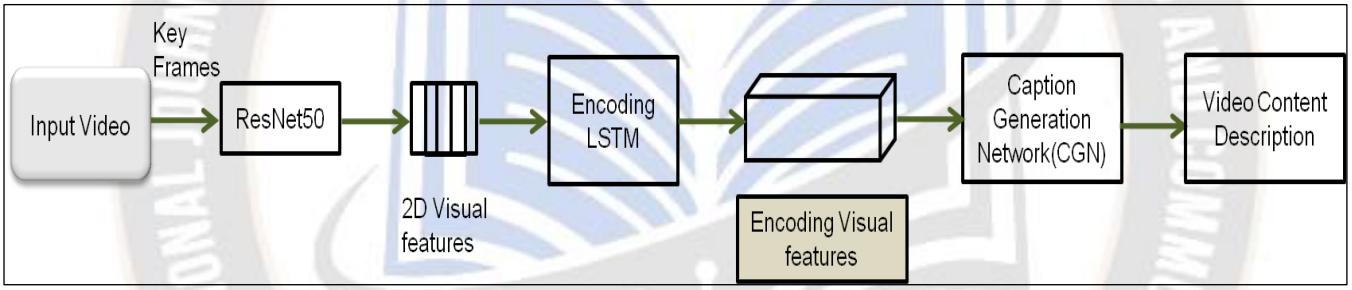

Figure 3: Feature Extraction using ResNet-50

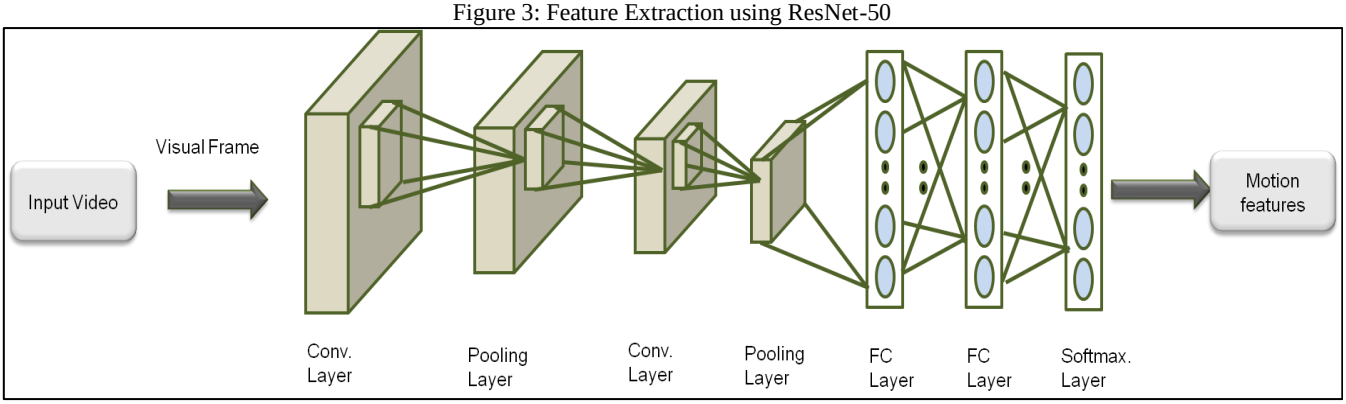

Figure 4: C3D feature extraction

B. ConvNet-LSTM

For extracting spatial information and learning temporal models, ConvNet-LSTM-based approaches use both the CNNs and the LSTMs. After 8 epochs, the Adam optimizer is employed with the initial learning rate set at 104 and reduced to 105 after 8 epochs for the ConvNet-LSTM and Two-Stream models (Pang et al. 2020). Adding a recurrent layer to a 2D ConvNet-LSTM model makes it more suitable for dynamic expression, since it introduces temporal information (Wang et al. 2022). Since there are 1024 hidden units, they use an LSTM layer with batch normalization. The development and assessment of activity proposals rely on frame-level elements like C3D.

C. TwoStream Networks

Because it only gets one kind of input, a single stream CNN is unable to comprehend the spatial and temporal aspects of human activities at the same time. Xiong et al. 2020 employed a spatial and temporal stream-based transferrable two-stream CNN architecture. To enhance action recognition, the recovered spatial and temporal information is concurrently analyzed by a two stream CNN structure. Using CNN to represent a spatial cue is easier and more economical in the two-stream architecture since the displacement values correspond to the moving scene points at the similar spatial location in many frames at the same temporal interval. Just a few frames in length are used to represent temporal cues because of the limited amount of time available. The two-stream CNN is employed to extract the spatial and short-term features of video frames.

D. C3D

C3D is a commonly used 3D CNN for extracting video's motion characteristics (Li et al. 2019). Fine motion information between successive frames is captured by this method by modelling a video in 3D spatiotemporal cuboid, this local motion

information is collected and maintained in higher-level representations. The identification of video events has been a hot topic in the scientific community for the last several decades. C2D (Convolution 2-dimensional) and other machine learning approaches have been employed, however C3D (Convolution 3-dimensional) has not been used for this purpose. Deep convolution networks are built to emphasize different video events by using C3D (Convolution 3-dimensional) to fully

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

leverage spatiotemporal relationships (Chen et al. 2020). C3D multimodal feature extraction is used by Peng et al. 2021 for video captioning. Figure 4 depicts the C3D feature extraction.

E. ImageNet Shuffle and Motion Boundary Histogram

The global video-level features, such as ImageNet Shuffle and motion boundary histogram (MBH), are used for untrimmed

video classification task (Daune et al. 2022). Specifically, for discrete semantics, the models are trained on ImageNet Shuffle

features for detecting both static and dynamic visual concepts. Training models on large-scale ImageNet shuffles based on encoding is beneficial for event detection. Febin et al. 2020 extracted motion boundary histogram features for violence detection in video surveillance. MBH is more robust to camera motion than optical flow, and thus more discriminative for action recognition.

F. Inflated 3D Convnet

Inflated 3D kernels may be seen as the orderly combination of 2D kernels. The video sequence training technique may teach the inflated 3D ConvNets abstract spatio-temporal representation. Clutters and confusing backdrops might have a negative impact on existing approaches, such as RGB image or optical flow. In order to deal with this problem, Wu et al. 2021 suggested a new Pose-Guided Inflated 3D ConvNet architecture. Based on the reuse of 2D architecture, an inflating 3D convnet achieved astounding results (Huang et al. 2020) [123].

TABLE 6: COMPARISON OF VIDEO INTERPRETATION FEATURES

Features	Description	Advantages	Limitations
Two Stream Network	Novel two-stream network consists of a "uniform sampling stream (USS) and an action pooling stream (APS)" to extract visual features (Yu et al. 2019) [107]	It is potential of capturing both global and action-local aspects of videos, and it's useful for comprehending lengthy, uncut footage.	The expressive potential of the two-stream characteristics is affected by the attentive model's reasoning ability.
	Two stream models extracts spatial and temporal features (Zhu et al. 2018) [109]	Effective for action recognition when the training dataset size is limited	False label assignment problem
ResNet 50	Multi-modal stochastic RNNs networks with ResNet 50 based feature extraction (Song et al. 2018) [38]	Efficiently extracts video representative features and outperforms GoogleNet and VGGNet	Higher training time for deeper network
ConvNet-LSTM	In this model, spatial information is extracted via ConvNets, and temporal patterns are modeled via LSTMs (Zalluhoglu and Ikizler-Cinbis 2020) [111]	Well suited for variable-length sequence prediction in videos	Relatively less successful than 3D convent
C3D	C3D network extracts motion features from videos (Peng et al. 2021) [39]	Single stream C3D features represent long-term temporal structure of actions from sparsely sampled short segments of video.	Does not consider spatial features
ImageNet Shuffle	It includes bottom-up and top-down shuffle for reorganization of the ImageNet hierarchy (Daune et al. 2022)	Tackles the image imbalance and over-specific class problems	Video Story embedding features slightly outperforms the ImageNet Shuffle

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Motion Boundary Histogram	MBH encodes the gradients of optical flow for video representation (Febin et al. 2020)	Efficiently captures information from datasets which may have taken while the camera might be moving	Computational complexity
Inflated 3D convnet	I3D convnet approaches capture RGB video frame features (Huang et al. 2020) [48]	Beneficial for action recognition where depth data is not required	Does not suit for dense video captioning

VII. Deep Learning in Video Interpretation Types

A. Image Captioning

Image captioning, is the act of providing a textual description that best conveys the visual scene or images in videos. The overfitting issue affected most contemporary image captioning methods in remote sensing, which failed to leverage semantic information in images. Shen et al. 2020 [110] suggested a Two-stage Multi-task Learning Model based on Variational Autoencoder and Reinforcement Learning for image captioning. Gupta and Jalal 2020 [112] suggested a model that combines a deep CNN and LSTM to improve image captioning accuracy by combining text information in an image with visual data. Chu et al. 2020 [113] proposed a combined model that can do automated image captioning using ResNet50 and LSTM along with soft attention.

B. Dense Image Description

Image Captioning creates simply a single description of the image. But it is less instructive. Hence, describing various Regions of Interest from a single image, also known as dense captioning is developing. The dense image captioning describes numerous parts of the image that include objects and some interactions between them. Dense description of different image areas is recognized as a superior understanding of the visual information. Particularly, the produced captions are able to give more fine-grained semantic information for image areas, which further allows complicated reasoning on visual context. Hence, dense captioning job may be applied in visual question answering (Liu et al. 2021) [98]. Therefore, it is seen as a more informative method for describing visuals, but also a more complex one. Kim et al. 2021 developed multi-task triple-stream network, a unique dense image captioning model which tries to create numerous captions with regard to relational data between objects in a visual picture. Duan et al. 2022 offer a Position-Aware Transformer (PAT) framework that captures static and regional visual characteristics and integrate these characteristics by including spatial information matched to each visual feature for dense captioning. Zhao et al. 2020 [105] introduced a unique Cross-scale Fusion with Global Attribute model (CSGA) that allows the dense caption model to execute regular end-to-end activation without mutual interference.

C. Activity Recognition and localization

In video interpretation and surveillance, recognizing human actions from videos has been a major challenge (Roy et al. 2021) [31]. Activity recognition is a popular study area right now. Recognition of human actions, e motions and gestures, are all included in activity recognition. The area of activity recognition has necessitated the development of several deep learning algorithms in the past. Previous techniques to video action recognition have always used the same principles as those used in image recognition. Human acts, on the other hand, are dynamic and constantly shifting, with a variety of target objects and diverse appearances in different contexts. Due to the enormous dimensions of the video data and the chaotic backdrop, human action detection from the wild films is still a difficult task. Video-based action recognition currently relies on two kinds of inputs: RGB pictures and the accompanying optical flow fields (Shi et al. 2022) [49]. For action recognition, Wang et al. 2018 [88] employed CNN, LSTM units, and a temporal-wise attention model. The action recognition task has been suggested to employ a variety of local space-time

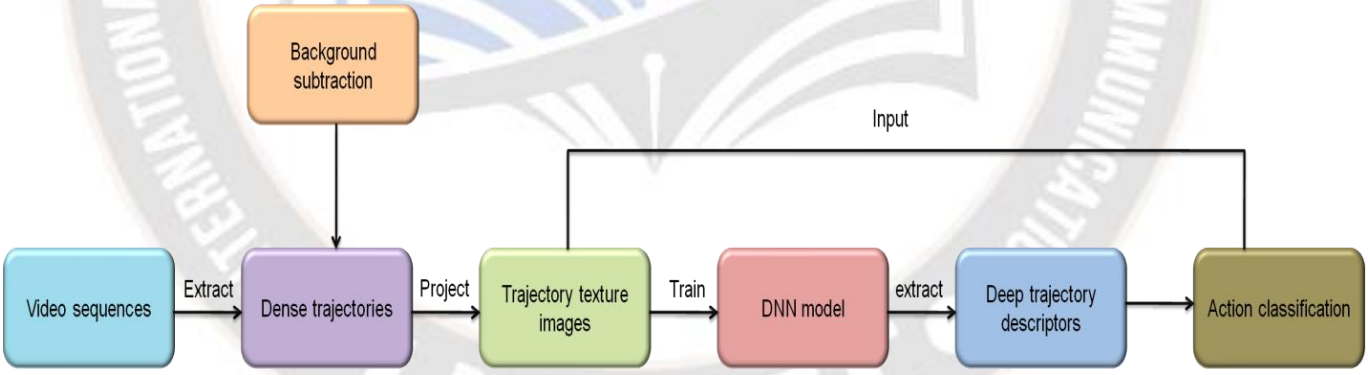

Figure 5: Activity Recognition Process

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

representations. A few examples are space-time interest points (STIPs) and dense trajectories. Figure 4 illustrates the use of dense trajectory characteristics for activity identification. Both STIPs and dense trajectories are plagued by the problem of overemphasizing non-static portions of the video while neglecting or ignoring static elements. Figure 5: Activity Recognition Process

Computer vision researchers are becoming interested in action localization, which is the process of locating distinct action sequences in videos (Zhao et al. 2020) [105]. Video label is called "weak" and may be used to develop models that can identify and locate activities in continuous videos (Luo et al. 2020) [102] (Zhai et al. 2020) [104]. Recent years have seen a flurry of research on the topic of temporally locating actions in untrimmed recordings (Min et al. 2020) [51]. For the newly suggested job of Temporal Activity Localization through Language (TALL) in video, fine-grained knowledge of the video material is required; unfortunately, most of the previous efforts fail to address this issue. The new TALL technique we provide in this research develops a hierarchical visual-textual graph to describe interactions between objects and words, as well as between objects themselves, to simultaneously grasp the video's content and the language used in the movie (Chen et al. 2020) [63].

D. Video Summarization

For an efficient browsing experience, video summarization seeks to provide a concise summation while still providing all of the relevant information. The perfect video summary is one that can deliver the most information in the least amount of time possible (Ji et al. 2019) [53]. Video indexing, video retrieval, and event recognition are just a few of the numerous practical uses for it. Storyboards and video skims are the two most used methods of video summary. A video skim is made up of a series of sample video segments known as key-shots, while a storyboard is based on a collection of keyframes. Video skim is the subject of this study. However, by choosing one or more keyframes from each key-shot, it may be easily transformed into a storyboard. Encoder-decoder video summary is shown in Figure 6. For the summarization of surveillance films acquired in IoT environments, Muhammed et al. 2019 employed a deep CNN architecture with hierarchical weighted fusion.

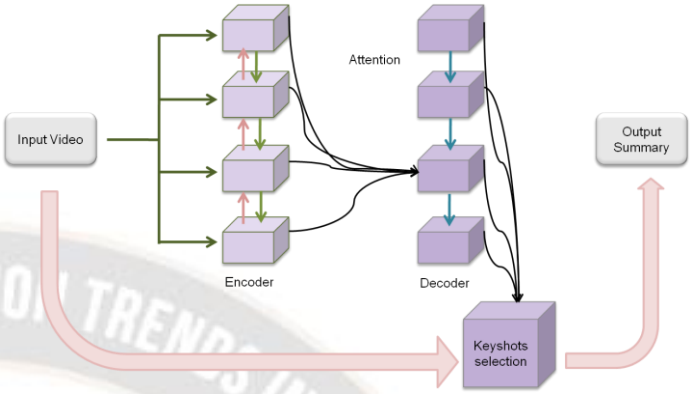

Figure 6: Video Summarization using encoder-decoder model

E. Video Captioning or Description

Template-based language models and sequence learning models dominate current techniques to video captioning. Before generating any sentences, the first step is to design a collection of language templates that adhere to specified grammatical norms. The produced sentences are constrained to a fixed syntactical structure since this technique heavily relies on predetermined templates and easily recognized phrases from films. Video footage may be immediately translated into a sentence using sequence learning models, which are extensively utilised in the machine translation sector (Xu et al. 2017) [94]. An encoder by CNNs with RNNs reads the entire video sequence and produces the video representation, and a decoder by RNNs then generates a natural sentence depending on syntactical structures learned from training data. This kind of network architecture is common. In video description, the technique of "temporal attention" has been frequently utilised to selectively concentrate on key frames. As a result, many current systems relying on temporal attention suffer from issues such as incorrect recognition or a lack of information. In light of recent successes in picture description utilizing spatial attention, Tu et al. 2017 suggest a spatial temporal attention (STAT) strategy to deal with these issues. This model is shown in Figure 7. Video Captioning using spatio-temporal attention

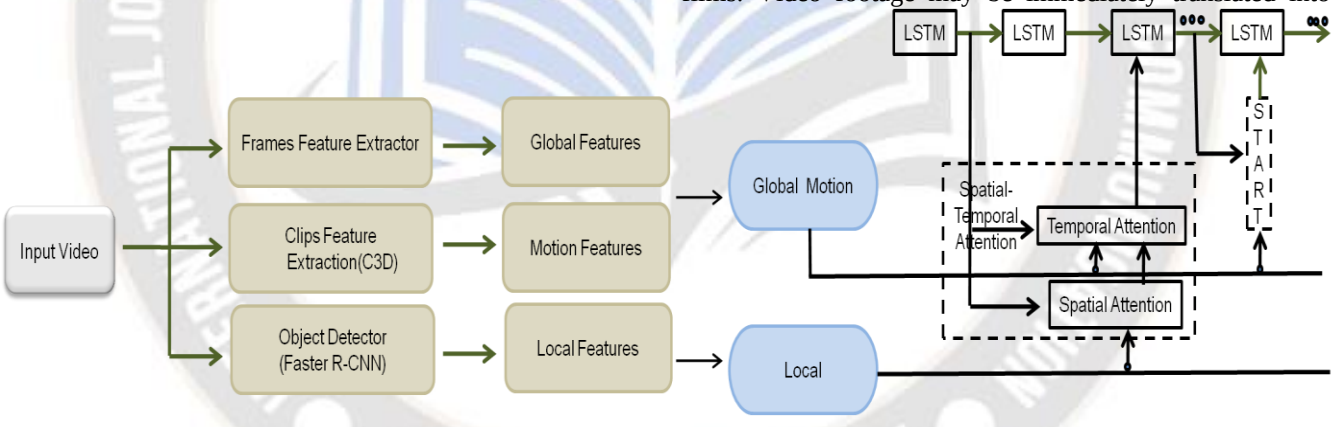

Figure 7: Video Captioning using spatio-temporal attention model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

model the procedure of the video description system is described below:

Input Data: Videos that have had explanations demanded are sent to the system for interpretation.

Scene Segmentation and Key Frame Extraction: The video is divided into a series of sections with varied lengths of time. To keep the scene's resolution adequate for creating the characterization, key frames are collected.

Feature Extraction: The significant visual features of the key video frames are extracted in the encoding stage.

Generating Video Description: In the decoding step, the model processes the visual properties of key frames to provide descriptions that best describe the scenario in the video. Any writing in the key frames, persons with ID (to manage reappearances), gender, mood, hair color, age, items, and surroundings are all included in the description.

F. Dense Video Captioning

The majority of video captioning research has concentrated on creating a single caption phrase for short recordings. A single phrase, on the other hand, is insufficient to comprehend or express multiple elements in extended films. We humans learn from films like this by paying attention to the subtitles. As a result, difficult projects like dense video captioning, which aims to describe all identified contents inside a lengthy movie with many natural language phrases at the same time, have gotten a lot of attention (Lin et al. 2018) [114]. Dense Video Captioning primarily entails two tasks: event detection, which identifies all occurrences in a brief video, and caption generation, which uses natural language phrases to explain the event suggestions (Fujita et al. 2020) [116]. For dense video captioning, Chang et al. 2022 [116] presented an event-centric multi-modal fusion technique. For dense video captioning, Zhang et al. 2020[118] presented a graph-based partition-and-summarization (GPaS) methodology.

VIII. Common Flow of Video Interpretation by Deep Learning

Deep learning architectures for encoding and decoding stages in video interpretation are explained below.

A. CNN - RNN Video Interpretation:

In this model, convolution architectures are used for visual encoding and recurrent structures are used for decoding (Khamparia et al. 2020) [118] (Emon et al. 2020) [120]. This is the most common architecture employed in deep learning-based video description methods. Garg et al. 2021[119] suggested an Encoder-Decoder approach where VGG19 CNN is working as Encoder; LSTM is working as Decoder to generate the text interpretation of video frames.

B. RNN - RNN Video Interpretation:

Here recurrent networks are used for both encoding and decoding stages (Feng et al.) [121]. Kumar and Verma 2021[124] described RNN-LSTM for caption generation for image or video frames.

C. Deep reinforcement networks:

They are a very recent topic of video description study. To improve the quality of the produced phrases for video frames, Reinforcement Learning is used (Hua et al. 2022) [123]. Gradient decay across layers may be caused by sigmoid and hyperbolic activation functions in LSTM and gated recurrent unit (GRU) based models employed in recent work on video summarization. Furthermore, due of the entanglement of neurons on RNN, analyzing and building network models is challenging. To address these challenges, Yaliniz and Ikizler-Cinbis 2021 [122] suggested an approach for unsupervised video summarizing that combines deep reinforcement learning and machine learning.

IX. Evaluation Metrics

It is impossible to compare the total quality of all sorts of models against the ground truth using any other evaluation criteria. A model's video captioning may be judged on how near it is to human annotation by evaluating metrics. The most significant metrics— "Bilingual Evaluation Understudy (BLEU), Consensus-based Image Description Evaluation (CIDEr), Metric for Evaluation of Translation with Explicit ORder-ing (METEOR), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)", and so on—have been fully discussed in this part.

A. BLEU

BLEU is a machine translation evaluation method. The range of values [0, 1] is obtained by calculating the cooccurrence word frequency of two sentences. The higher the score, the closer the machine translation is to the original human translation. Because the computation is so rapid and so near to human judgement, it has an advantage. The drawback is that no consideration is given to grammar, synonyms, or other related semantics, and the accuracy of short statements is easily abused. N-gram matching rules may be used in conjunction with the shortness penalty (BP) computation in BLEU (Deng et al. 2021) [58]. BELU can be calculated using equation 1.

$$\log \text{BLEU} = \min\left(1 - \frac{b_j}{b_x}, 0\right) + \sum_{q=1}^{Q} l_q log\, T_q \quad (1)$$

Where bj/bx refers to the ratio of the length of the reference corpus to the length of the candidate description, lq means the positive weight, and Tq denotes the geometric mean of the modified n-gram precisions.

B. METEOR

As BLEU has intrinsic flaws, METEOR is an algorithm to address them. In the range of [0, 1], it returns. To broaden the pool of possible synonyms and take the word form into account, it consults a knowledge base like WordNet. The harmonic mean of accuracy and recall of unigram matches between texts is used to calculate the results. The idea of chunks is used to evaluate the fluency of sentences. It is calculated using equation 2.
METEOR=S(1-T)

$$METEOR = S(1 - T) \quad (2)$$

Where T means penalty and S means the F-measure

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

C. ROUGE

Automated summaries and machine translations are evaluated using ROUGE, a computer programmer. It is similar to BLEU in that the co-occurrence frequency is calculated using n-grams. Because the recall ranges from 0 to 1, it is utilized as an indication rather than a metric. The most important thing to notice is how many times a cited phrase appears in the text. It is determined using equation 3.

$$\mathrm{ROUGE-N} = \frac{\sum_{U \in Jsum} \sum_{k \in U} X_m(k_q)}{\sum_{U \in Jsum} \sum_{k \in U} X_m(k_q)} \quad (3)$$

Where q means the n-gram length, *kq*, and *Xm(kq)* denote the maximum number of n-grams that are available in candidate and ground truth summaries respectively and *Jsum* means the reference summaries.

D. CIDEr

Unlike the other three assessment measures for machine translation, CIDEr uses an algorithm for assessing picture captions. When using CIDEr, every phrase is treated as a separate document by the software. The n-gram mass is calculated using the TF-IDF algorithm. Using the cosine distance, its similarity may be determined. As part of the n-grams computation, the accuracy and recall are taken into account (i.e., the higher and the better). TF-IDF weights for various n-grams are different because less information is included in more frequent n-grams in the corpus. Due to the importance of capturing critical data, non-keyword reduction is required for assessment. It is determined using equation 4.

$$\mathrm{CIDEr}_q = (x_w, U_w) = \frac{1}{b} \sum_o \frac{k^q(x_w).k^q(U_{wo})}{\|k^q(x_w)\|.\|k^q(U_{wo})\|} \quad (4)$$

E. Rank Based Measures

The video captions are ranked according to how well they match the video or image query in terms of visual feature similarity. The caption rating is used to assess the performance. R@K (Recall at K) is one of the rank-based measures in video description (Dong et al. 2018) [55]. For each test image or video, R@K calculates the proportion of valid descriptions found in top-K recovered captions. Higher R@K results in improved video description performance.

F. WMD

Word embeddings, which are vector representations of words learned from text corpora, are used in Word Mover's Distance (WMD). Dissimilarity between two papers is measured using WMD distance. Even though the words in two captions are spelled differently, their semantic meanings may be the same. If many captions have similar properties, objects, and relations but convey radically distinct meanings, this is possible. This issue was addressed by the use of WMDs (Fujita et al. 2020)[116].

G. Semantic Textual Similarity (STS)

This metric compares the provided description to one of the ground truth explanations in terms of semantic similarity (Smeaton et al. 2019) [61]. STS determines the inter-caption semantic similarity.

H. Direct Assessment

In 2018, Graham et.al., [59]. released Direct Assessment, a technique for manually evaluating the quality of automated video captions. Crowdsourcing the quality of a caption's description of a video adds human judgement to the review process. In the major Machine Translation benchmark assessments, DA is currently the official technique of ranking. To evaluate video captions, DA presents a film and a single caption to human evaluators. Assessors are asked to give a score between 0 and 100 based on how effectively the video's captions capture the action.

I. PR 15

Percentage miss (Pmiss) and the rate of false alarms (RFA) were used to calculate the decision threshold, which was Pmiss at RFA = 0.15. (PR.15). False Alarm (FA) indicates that a captioning instance in the system output has no connectivity to the reference, while Missed Detection (MD) indicates that a captioning instance in the reference has no connectivity to the system output. PR 15 is computed using equation 5 and 6.

$$P_{miss}(p) = \frac{Q_{BS}(p)}{Q_{TrueInstance}} \quad (5)$$

$$R_{FA}(p) = \frac{Q_{SZ}(P)}{Video\ Duration} \quad (6)$$

Where QBS(p) refers to the quantity of missed detections at the threshold p, QSZ(p) means the number of false alarms, and QTrueInstance refers to the number of reference examples labeled in the video sequence.

J. nAUDC

The normalized Area Under the Detection Error Tradeoff (DET) curve (nAUDC) metric focus more on recall rather than precision, and the alignment is strictly for short activities and loosely for long activities (Godil et al. 2021) [62].

K. Human Evaluations

Human evaluations involve manual judgment of reliability regarding video interpretations made by automatic models. Relevance and Grammar Correctness are two examples of metrics that may be used to shape human judgments.

X. BENCHMARK RESULTS

Deep learning-based video interpretation consists of two main events like visual content extraction and its representation using dynamic feature vector and text generation from feature vectors. Table 4 shows the performance of various deep learning models on benchmark datasets in recent years. Some of the popular datasets in the recent years include “MSRVTT, M-VAD, MPII-MD, MSVD, TACoS-Multi-Level, ActivityNet”, and so on. The evaluation metrics used for the comparison includes BeLu, CIDER, ROUGE, and METEOR. MSVD is the most widely used dataset, and it may perform the best in many video captioning algorithms. The quantity of natural language phrases supplied each video clip in a dataset has a big impact on the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

captioning algorithms' reliability. When contrasted to other datasets, MPII-MD performs poorly.

XI. Future and Emerging Trends

Deep learning (DL) architectures have recently evolved and gotten more complex in terms of architecture and processing, in order to keep up with the newest advances in video processing applications. Still, in terms of processing power, DL architectures need to be more prominent. Each computer vision application has its own unique features. Some computer vision applications, for example, need pixel-level annotations, while others want object-level annotations, and still others demand scene-level annotations. Deeper DL architectures must be constructed by employing additional layers to enhance the video analysis performance. Hierarchical learning of features (learning variant features from variant layers) is one of the useful solutions in this field. Multidimensional features other than conventional features must be employed in model training. Each computer vision application has its own unique features. Some computer vision applications, for example, need pixel-level annotations, while others want object-level annotations, and still others demand scene-level annotations. To improve performance and achieve state-of-the-art results, contemporary improvements in video captioning include "reinforcement learning and teacher-recommended object relational learning". The keyframe's capacity to interpret complex behavior's aids researchers in developing more efficient activity identification systems (Dang et al. 2020) [106]. An important emerging trend in video captioning is the application of transfer learning in video analytics. For video preprocessing, compression, analysis, and interpretation, new highly powerful algorithms must be developed. The algorithms must be able to detect and remember significant occurrences, as well as alert anything that would be considered strange. They must be able to comprehend each case based on the participants and circumstances, as well as synthesize findings, draw inferences, and make forecasts. On the other hand, clearer and explicit privacy and data security protocols must be created and executed with narrow tolerance in activity recognition. An essential expansion of Video Captioning, Stylized Captioning integrates the notion of style-transfer (or feature switching) from the vision domain with a technique of producing captions from visual input. For example, in this area of study, the development of robust models may improve a model's capacity to create better captions by using vision-inspired techniques like representation learning and disentanglement. It's been a decade since the task of visual question answering (VQA) has multiplied by the development of attention frameworks that promote improved comprehension abilities in modern VQA systems. Explainability-driven frameworks like hierarchical and graph-based attention have been quite useful in this study. Free-form question-answering is becoming more common in VQA due to recent advances in different language generating tasks, whereas MCQ responses are becoming less common. Co-attention based VQA has emerged as the most often used attention framework because of

XII. Summary and Discussion

A brief history of video interpretation is provided in the survey's introduction. Examining in a replay version to identify any activity or scene might be tiresome work since there would only be motion for a brief period of time in the video. Analyzing the video by this manual technique would take a long time, and it would be hard to always describe the video accurately. Hence, there is a need for an automatic video interpretation approach. But developing these models and analysis has certain challenges. Some of the difficulties arise from the wide variety of videos. A model's video description performance may suffer if there are several activities in a video, but only part of the activities are represented by captions. Some elements of dataset videos, such as the similarity of motions, cluttering backdrop and views, lighting change and occlusion, may be seen to have limits. In the absence of a huge video dataset with many different activity classifications and massive numbers of films and subtitles, this poses a significant problem. Longer videos also present additional challenges, as most action features, such as trajectory and C3D, shall only encode short-term actions because of the reliance on video segment lengths. Different video feature extractors have difficulty dealing with sudden changes in the scene. The visual encoding process is now simplified by expressing movies or frames as a whole. Further attention models may be needed to concentrate on the video's most important spatial and temporal aspects. Instead of untangling the graphical depiction from the temporal model and the temporal model itself from language, more focus should be placed on constructing stronger temporal modeling structures. Deep learning models exhibit promising results in video interpretation.

XIII. Conclusion

We have provided a comprehensive literature review for video interpretation research, ranging from traditional methodologies to more modern statistical and deep learning-based techniques. The benchmark datasets for video description models were investigated in this survey research. We also looked at the presently available assessment measures for evaluating the video captions produced, emphasizing the necessity for particular and defined datasets and evaluation criteria to improve performance. Single stream features like C3D only considers temporal features or spatial features which is not efficient for video interpretation. Two stream features are beneficial for video understanding compared to single stream. From the analysis, it is observed that ResNet 50, C3D features are better than VGGNet and GoogleNet, for video interpretation. MSVD is the most widely used dataset compared to other datasets like MSRVTT, Activitynet, MPII-MD, TACoS and so on, and it may perform the best in many video captioning algorithms. Finally, we provide some suggestions for future study approaches that are likely to further this field's frontiers of exploration. Dense video captioning, visual question answering, stylized video captioning, transfer learning in video analytics are all emerging trends in video interpretation.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to the Department of Technology Shivaji University Kolhapur for their continuous support throughout the submission and evaluation process of this work. Their assistance and guidance have been invaluable in shaping this research. Thank you for your unwavering commitment and encouragement.

REFERENCES

[1] Gao, L., Guo, Z., Zhang, H., Xu, X. and Shen, H.T., 2017. Video captioning with attention-based LSTM and semantic consistency. IEEE Transactions on Multimedia, 19(9), pp.2045-2055.

[2] Ramanishka, V., Das, A., Park, D.H., Venugopalan, S., Hendricks, L.A., Rohrbach, M. and Saenko, K., 2016, October. Multimodal video description. In Proceedings of the 24th ACM international conference on Multimedia (pp. 1092-1096).

[3] Yuksel, B.F., Kim, S.J., Jin, S.J., Lee, J.J., Fazli, P., Mathur, U., Bisht, V., Yoon, I., Siu, Y.T. and Miele, J.A., 2020, April. Increasing video accessibility for visually impaired users with human-in-the-loop machine learning. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-9).

[4] Zanfir, M., Marinoiu, E. and Sminchisescu, C., 2016, November. Spatio-temporal attention models for grounded video captioning. In asian conference on computer vision (pp. 104-119). Springer, Cham.

[5] Bin, Y., Yang, Y., Shen, F., Xie, N., Shen, H.T. and Li, X., 2018. Describing video with attention-based bidirectional LSTM. IEEE transactions on cybernetics, 49(7), pp.2631-2641.

[6] Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A. and Schiele, B., 2017. Movie description. International Journal of Computer Vision, 123(1), pp.94-120.

[7] Gao, L., Wang, X., Song, J. and Liu, Y., 2020. Fused GRU with semantic-temporal attention for video captioning. Neurocomputing, 395, pp.222-228.

[8] Islam, S., Dash, A., Seum, A., Raj, A.H., Hossain, T. and Shah, F.M., 2021. Exploring video captioning techniques: A comprehensive survey on deep learning methods. SN Computer Science, 2(2), pp.1-28.

[9] Aafaq, N., Mian, A., Liu, W., Gilani, S.Z. and Shah, M., 2019. Video description: A survey of methods, datasets, and evaluation metrics. ACM Computing Surveys (CSUR), 52(6), pp.1-37.

[10] Rafiq, M., Rafiq, G. and Choi, G.S., 2021. Video Description: Datasets & Evaluation Metrics. IEEE Access, 9, pp.121665-121685.

[11] Suresha, M., Kuppa, S. and Raghukumar, D.S., 2020. A study on deep learning spatiotemporal models and feature extraction techniques for video understanding. International Journal of Multimedia Information Retrieval, 9(2), pp.81-101.

[12] Singh, T. and Vishwakarma, D.K., 2019. Human activity recognition in video benchmarks: A survey. In Advances in Signal Processing and Communication (pp. 247-259). Springer, Singapore.

[13] Dave, J. and Padmavathi, S., 2022. Hierarchical Language Modeling for Dense Video Captioning. In Inventive Computation and Information Technologies (pp. 421-431). Springer, Singapore.

[14] Yu, Z. and Han, N., 2021. Accelerated masked transformer for dense video captioning. Neurocomputing, 445, pp.72-80.

[15] Xu, Y., Han, Y., Hong, R. and Tian, Q., 2018. Sequential video VLAD: Training the aggregation locally and temporally. IEEE Transactions on Image Processing, 27(10), pp.4933-4944.

[16] Ding, S., Qu, S., Xi, Y. and Wan, S., 2019. A long video caption generation algorithm for big video data retrieval. Future Generation Computer Systems, 93, pp.583-595.

[17] Droste, R., Jiao, J. and Noble, J.A., 2020, August. Unified image and video saliency modeling. In European Conference on Computer Vision (pp. 419-435). Springer, Cham.

[18] Dilawari, A., Khan, M.U.G., Farooq, A., Rehman, Z.U., Rho, S. and Mehmood, I., 2018. Natural language description of video streams using task-specific feature encoding. IEEE Access, 6, pp.16639-16645.

[19] Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M. and Gan, C., 2020. Dense regression network for video grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10287-10296).

[20] Pan, Z. and Li, C., 2020. Robust basketball sports recognition by leveraging motion block estimation. Signal Processing: Image Communication, 83, p.115784.

[21] Jaouedi, N., Boujnah, N. and Bouhlel, M.S., 2020. A new hybrid deep learning model for human action recognition. Journal of King Saud University-Computer and Information Sciences, 32(4), pp.447-453.

[22] Tu, Y., Zhou, C., Guo, J., Gao, S. and Yu, Z., 2021. Enhancing the alignment between target words and corresponding frames for video captioning. Pattern Recognition, 111, p.107702.

[23] Saleem, S., Dilawari, A., Khan, U.G., Iqbal, R., Wan, S. and Umer, T., 2019. Stateful human-centered visual captioning system to aid video surveillance. Computers & Electrical Engineering, 78, pp.108-119.

[24] Sun, B., Wu, Y., Zhao, K., He, J., Yu, L., Yan, H. and Luo, A., 2021. Student Class Behavior Dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. Neural Computing and Applications, 33(14), pp.8335-8354.

[25] Shin, A., Ishii, M. and Narihira, T., 2022. Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision. International Journal of Computer Vision, pp.1-20.

[26] Bhowmik, A., Kumar, S. and Bhat, N., 2021. Evolution of automatic visual description techniques-a methodological survey. Multimedia Tools and Applications, 80(18), pp.28015-28059.

[27] Poorgholi, S., Kayhan, O.S. and Gemert, J.C.V., 2021, January. t-eva: Time-efficient t-sne video annotation. In International Conference on Pattern Recognition (pp. 153-169). Springer, Cham.

[28] Zhang, C., Lin, Y., Zhu, L., Liu, A., Zhang, Z. and Huang, F., 2019. CNN-VWII: An efficient approach for large-scale video retrieval by image queries. Pattern Recognition Letters, 123, pp.82-88.

[29] Gao, L., Li, T., Song, J., Zhao, Z. and Shen, H.T., 2020. Play and rewind: Context-aware video temporal action proposals. Pattern Recognition, 107, p.107477.

[30] Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G. and Fei-Fei, L., 2018. Every moment counts: Dense detailed labeling of actions in complex videos. International Journal of Computer Vision, 126(2), pp.375-389.

[31] Roy, A., Banerjee, B., Hussain, A. and Poria, S., 2021. Discriminative Dictionary Design for Action Classification in Still Images and Videos. Cognitive Computation, 13(3), pp.698-708.

[32] Zuo, Z., Yang, L., Liu, Y., Chao, F., Song, R. and Qu, Y., 2019. Histogram of fuzzy local spatio-temporal descriptors for video action recognition. IEEE Transactions on Industrial Informatics, 16(6), pp.4059-4067.

[33] Mou, L. and Zhu, X.X., 2016, July. Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis. In 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1823-1826). IEEE.

[34] Choi, S.H., Jo, S.Y. and Jung, S.H., 2021. Component based comparative analysis of each module in image captioning. ICT Express, 7(1), pp.121-125.

[35] Suresh, A.J. and Visumathi, J., 2020. Inception ResNet deep transfer learning model for human action recognition using LSTM. Materials Today: Proceedings.

[36] Hammad, M., Hammad, M. and Elshenawy, M., 2020, October. Characterizing the impact of using features extracted from pre-trained models on the quality of video captioning sequence-to-sequence models. In International Conference on Pattern Recognition and Artificial Intelligence (pp. 238-250). Springer, Cham.

[37] Lee, S. and Kim, I., 2018. Multimodal feature learning for video captioning. Mathematical Problems in Engineering, 2018.

[38] Song, J., Guo, Y., Gao, L., Li, X., Hanjalic, A. and Shen, H.T., 2018. From deterministic to generative: Multimodal stochastic RNNs for video captioning. IEEE transactions on neural networks and learning systems, 30(10), pp.3047-3058.

[39] Peng, Y., Wang, C., Pei, Y. and Li, Y., 2021. Video captioning with global and local text attention. The Visual Computer, pp.1-12.

[40] Li, X., Zhou, Z., Chen, L. and Gao, L., 2019. Residual attention-based LSTM for video captioning. World Wide Web, 22(2), pp.621-636.

[41] Chen, S., Jiang, W., Liu, W. and Jiang, Y.G., 2020, August. Learning modality interaction for temporal sentence localization and event captioning in videos. In European Conference on Computer Vision (pp. 333-351). Springer, Cham.

[42] Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W. and Zhang, W., 2022. FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos. arXiv preprint arXiv:2203.09463.

[43] Pang, B., Zha, K., Zhang, Y. and Lu, C., 2020, April. Further understanding videos through adverbs: A new video task. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 11823-11830).

[44] Xiong, Q., Zhang, J., Wang, P., Liu, D. and Gao, R.X., 2020. Transferable two-stream convolutional neural network for human action recognition. Journal of Manufacturing Systems, 56, pp.605-614.

[45] Duane, A. and Jónsson, B.Þ., 2022. ViRMA: Virtual Reality Multimedia Analytics at Video Browser Showdown 2022. In International Conference on Multimedia Modeling (pp. 580-585). Springer, Cham.

[46] Wu, Q., Zhu, A., Cui, R., Wang, T., Hu, F., Bao, Y. and Snoussi, H., 2021. Pose-Guided Inflated 3D ConvNet for action recognition in videos. Signal Processing: Image Communication, 91, p.116098.

[47] Huang, Y., Guo, Y. and Gao, C., 2020. Efficient parallel inflated 3D convolution architecture for action recognition. IEEE Access, 8, pp.45753-45765.

[48] Febin, I.P., Jayasree, K. and Joy, P.T., 2020. Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. Pattern Analysis and Applications, 23(2), pp.611-623.

[49] Shi, L., Zhang, Y., Cheng, J. and Lu, H., 2022. Action recognition via pose-based graph convolutional networks with intermediate dense supervision. Pattern Recognition, 121, p.108170.

[50] Wang, L., Xu, Y., Cheng, J., Xia, H., Yin, J. and Wu, J., 2018. Human action recognition by learning spatio-temporal features with deep neural networks. IEEE access, 6, pp.17913-17922.

[51] Min, K. and Corso, J.J., 2020, August. Adversarial background-aware loss for weakly-supervised temporal activity localization. In European conference on computer vision (pp. 283-299). Springer, Cham.

[52] Chen, S. and Jiang, Y.G., 2020, August. Hierarchical visual-textual graph for temporal activity localization via language. In European Conference on Computer Vision (pp. 601-618). Springer, Cham.

[53] Ji, Z., Xiong, K., Pang, Y. and Li, X., 2019. Video summarization with attention-based encoder–decoder networks. IEEE Transactions on Circuits and Systems for Video Technology, 30(6), pp.1709-1717.

[54] Muhammad, K., Hussain, T., Tanveer, M., Sannino, G. and de Albuquerque, V.H.C., 2019. Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. IEEE Internet of Things Journal, 7(5), pp.4455-4463.

[55] Dong, J., Li, X. and Snoek, C.G., 2018. Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia, 20(12), pp.3377-3388.

[56] Xu, J., Yao, T., Zhang, Y. and Mei, T., 2017, October. Learning multimodal attention LSTM networks for video captioning. In Proceedings of the 25th ACM international conference on Multimedia (pp. 537-545).

[57] Tu, Y., Zhang, X., Liu, B. and Yan, C., 2017, October. Video description with spatial-temporal attention. In Proceedings of the 25th ACM international conference on Multimedia (pp. 1014-1022).

[58] Deng, C., Chen, S., Chen, D., He, Y. and Wu, Q., 2021. Sketch, Ground, and Refine: Top-Down Dense Video Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 234-243).

[59] Graham, Y., Awad, G. and Smeaton, A., 2018. Evaluation of automatic video captioning using direct assessment. PloS one, 13(9), p.e0202789.

[60] Fujita, S., Hirao, T., Kamigaito, H., Okumura, M. and Nagata, M., 2020, August. SODA: Story oriented dense video captioning evaluation framework. In European Conference on Computer Vision (pp. 517-531). Springer, Cham.

[61] Smeaton, A.F., Graham, Y., McGuinness, K., O'Connor, N.E., Quinn, S. and Arazo Sanchez, E., 2019, January. Exploring the Impact of Training Data Bias on Automatic Generation of Video Captions. In International Conference on Multimedia Modeling (pp. 178-190). Springer, Cham.

[62] Godil, A., Lee, Y., Fiscus, J., Delgado, A., Godard, E., Chocot, B., Diduch, L., Golden, J. and Zhang, J., 2021, January. 2020 Sequestered Data Evaluation for Known Activities in Extended Video: Summary and Results. In 2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW) (pp. 51-59). IEEE.

[63] Chen H, Li J, Hu X. Delving deeper into the decoder for video captioning. 2020. arXiv preprint arXiv:200105614

[64] J, Jia Y, Qi Y, et al. Video captioning using weak annotation. 2020 arXiv preprint arXiv:200901067

[65] Zhang X, Liu C, Chang F. Guidance module network for video captioning. 2020 arXiv preprint arXiv:201210930

[66] Perez-Martin J, Bustos B, Pérez J. Attentive visual semantic specialized network for video captioning. In: International Conference on Computer Vision 2020.

[67] Perez-Martin J, Bustos B, Perez J. Improving video captioning with temporal composition of a visual-syntactic embedding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021;3039–3049.

[68] Xiao H, Shi J. Video captioning with text-based dynamic attention and step-by-step learning. Pattern Recognition Letters. 2020

[69] Liu S, Ren Z, Yuan J. Sibnet: Sibling convolutional encoder for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020.

[70] Hou J, Jia Y, Qi Y, et al. Video captioning using weak annotation. 2020 arXiv preprint arXiv:200901067

[71] Iashin V, Rahtu E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. 2020. arXiv preprint arXiv:200508271

[72] Lei J, Wang L, Shen Y, Yu D, Berg TL, Bansal M. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. 2020. arXiv preprint arXiv:200505402

[73] Sur C. Sact: Self-aware multi-space feature composition transformer for multinomial attention for video captioning. 2020. arXiv preprint arXiv:200614262

[74] Iashin V, Rahtu E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. 2020. arXiv preprint arXiv:200508271

[75] Suin M, Rajagopalan A. An efficient framework for dense video captioning. In: AAAI, 2020;12039–12046.

[76] Hao X, Zhou F, Li X. Scene-edge gru for video caption. In: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, vol 1, 2020;1290–1295

[77] Wang T, Zheng H, Yu M, Tian Q, Hu H. Event-centric hierarchical representation for dense video captioning. IEEE Transactions on Circuits and Systems for Video Technology. 2020

[78] Xu J, Mei T, Yao T, Rui Y. Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;5288–5296.

[79] Zhang Z, Xu D, Ouyang W, Tan C (2019) Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. IEEE Transactions on Circuits and Systems for Video Technology

[80] Park JS, Rohrbach M, Darrell T, Rohrbach A. Adversarial inference for multi-sentence video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019;6598–6608

[81] Wang X, Wu J, Zhang D, Su Y, Wang WY. Learning to compose topic-aware mixture of experts for zero-shot video captioning. Proceedings of the AAAI Conference on Artifcial Intelligence. 2019b;33:8965–72.

[82] Zhang J, Peng Y. Hierarchical vision-language alignment for video captioning. In: International Conference on Multimedia Modeling, Springer, 2019a;42–54.

[83] Yan C, Tu Y, Wang X, Zhang Y, Hao X, Zhang Y, Dai Q. Stat: spatial-temporal attention mechanism for video captioning. IEEE transactions on multimedia 2019.

[84] Guo Y, Zhang J, Gao L. Exploiting long-term temporal dynamics for video captioning. World Wide Web. 2019;22(2):735–49

[85] Aafaq N, Akhtar N, Liu W, Gilani SZ, Mian A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019a;12487–12496.

[86] Chen J, Pan Y, Li Y, Yao T, Chao H, Mei T. Temporal deformable convolutional encoder-decoder networks for video captioning. Proceedings of the AAAI Conference on Artifcial Intelligence. 2019b;33:8167–74.

[87] Chen S, Jiang YG. Motion guided spatial attention for video captioning. Proceedings of the AAAI Conference on Artifcial Intelligence. 2019;33:8191–8.

[88] Wang B, Ma L, Zhang W, Liu W. Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018a;7622–7631.

[89] Wu X, Li G, Cao Q, Ji Q, Lin L. Interpretable video captioning via trajectory structured localization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018;6829–6837

[90] Wang J, Wang W, Huang Y, Wang L, Tan T. M3: Multimodal memory modelling for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018c;7512–7520.

[91] Xu N, Liu AA, Wong Y, Zhang Y, Nie W, Su Y, Kankanhalli M. Dual-stream recurrent neural network for video captioning. IEEE Transactions on Circuits and Systems for Video Technology. 2018;29(8):2482–93.

[92] B, Ma L, Zhang W, Liu W. Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018a;7622–763

[93] Yang Y, Zhou J, Ai J, Bin Y, Hanjalic A, Shen HT, Ji Y. Video captioning by adversarial lstm. IEEE Transactions on Image Processing. 2018;27(11):5600–11.

[94] Xu N, Liu AA, Wong Y, Zhang Y, Nie W, Su Y, Kankanhalli M. Dual-stream recurrent neural network for video captioning. IEEE Transactions on Circuits and Systems for Video Technology. 2018;29(8):2482–93.

[95] X, Gan C, de Melo G. Video captioning with multi-faceted attention. Transactions of the Association for Computational Linguistics. 2018;6:173–84.

[96] Wu X, Li G, Cao Q, Ji Q, Lin L. Interpretable video captioning via trajectory structured localization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018;6829–6837.

[97] Xu N, Liu AA, Wong Y, Zhang Y, Nie W, Su Y, Kankanhalli M. Dual-stream recurrent neural network for video captioning. IEEE Transactions on Circuits and Systems for Video Technology. 2018;29(8):2482–93

[98] Liu, A.A., Wang, Y., Xu, N., Liu, S. and Li, X., 2021. Scene-Graph-Guided message passing network for dense captioning. Pattern Recognition Letters, 145, pp.187-193.

[99] Kim, D.J., Oh, T.H., Choi, J. and Kweon, I.S., 2021. Dense relational image captioning via multi-task triple-stream networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[100] Zhao, D., Chang, Z. and Guo, S., 2020. Cross-scale fusion detection with global attribute for dense captioning. Neurocomputing, 373, pp.98-108.

[101] Duan, Y., Wang, Z., Wang, J., Wang, Y.K. and Lin, C.T., 2022. Position-Aware Image Captioning with Spatial Relation. Neurocomputing.

[102] Roy, C., Nourani, M., Honeycutt, D.R., Block, J.E., Rahman, T., Ragan, E.D., Ruozzi, N. and Gogate, V., 2021. Explainable activity recognition in videos: Lessons learned. Applied AI Letters, p.e59.

[103] Luo, Z., Guillory, D., Shi, B., Ke, W., Wan, F., Darrell, T. and Xu, H., 2020, August. Weakly-supervised action localization with expectation-maximization multi-instance learning. In European conference on computer vision (pp. 729-745). Springer, Cham.

[104] Zhai, Y., Wang, L., Tang, W., Zhang, Q., Yuan, J. and Hua, G., 2020, August. Two-stream consensus network for weakly-supervised temporal action localization. In European conference on computer vision (pp. 37-54). Springer, Cham.

[105] Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y. and Tian, Q., 2020, August. Bottom-up temporal action localization with mutual regularization. In European Conference on Computer Vision (pp. 539-555). Springer, Cham.

[106] Dang, L.M., Min, K., Wang, H., Piran, M.J., Lee, C.H. and Moon, H., 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition, 108, p.107561.

[107] Yu, T., Yu, J., Yu, Z. and Tao, D., 2019. Compositional attention networks with two-stream fusion for video question answering. IEEE Transactions on Image Processing, 29, pp.1204-1218.

[108] Sun, B., Wu, Y., Zhao, K., He, J., Yu, L., Yan, H. and Luo, A., 2021. Student Class Behavior Dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. Neural Computing and Applications, 33(14), pp.8335-8354.

[109] Zhu, Y., Lan, Z., Newsam, S. and Hauptmann, A., 2018, December. Hidden two-stream convolutional networks for action recognition. In Asian conference on computer vision (pp. 363-378). Springer, Cham.

[110] Shen, X., Liu, B., Zhou, Y., Zhao, J. and Liu, M., 2020. Remote sensing image captioning via Variational Autoencoder and Reinforcement Learning. Knowledge-Based Systems, 203, p.105920.

[111] Zalluhoglu, C. and Ikizler-Cinbis, N., 2020. Collective Sports: A multi-task dataset for collective activity recognition. Image and Vision Computing, 94, p.103870.

[112] Gupta, N. and Jalal, A.S., 2020. Integration of textual cues for fine-grained image captioning using deep CNN and LSTM. Neural Computing and Applications, 32(24), pp.17899-17908.

[113] Chu, Y., Yue, X., Yu, L., Sergei, M. and Wang, Z., 2020. Automatic image captioning based on ResNet50 and LSTM with soft attention. Wireless Communications and Mobile Computing, 2020.

[114] Lin, X., Jin, Q., Chen, S., Song, Y. and Zhao, Y., 2018, September. iMakeup: Makeup Instructional Video Dataset for Fine-Grained Dense Video Captioning. In Pacific Rim Conference on Multimedia (pp. 78-88). Springer, Cham.

[115] Fujita, S., Hirao, T., Kamigaito, H., Okumura, M. and Nagata, M., 2020, August. SODA: Story oriented dense video captioning evaluation framework. In European Conference on Computer Vision (pp. 517-531). Springer, Cham.

[116] Chang, Z., Zhao, D., Chen, H., Li, J. and Liu, P., 2022. Event-centric multi-modal fusion method for dense video captioning. Neural Networks, 146, pp.120-129.

[117] Zhang, Z., Xu, D., Ouyang, W. and Zhou, L., 2020. Dense video captioning using graph-based sentence summarization. IEEE Transactions on Multimedia, 23, pp.1799-1810.

[118] Khamparia, A., Pandey, B., Tiwari, S., Gupta, D., Khanna, A. and Rodrigues, J.J., 2020. An integrated hybrid CNN–RNN model for visual description and generation of captions. Circuits, Systems, and Signal Processing, 39(2), pp.776-788.

[119] Garg, K., Singh, V. and Tiwary, U.S., 2021, December. Textual Description Generation for Visual Content Using Neural Networks. In International Conference on Intelligent Human Computer Interaction (pp. 16-26). Springer, Cham.

[120] Emon, S.H., Annur, A.H.M., Xian, A.H., Sultana, K.M. and Shahriar, S.M., 2020, December. Automatic Video Summarization from Cricket Videos Using Deep Learning. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[121] Feng, X., He, X., Huang, R. and Liu, C., 2021, November. A Fully Dynamic Context Guided Reasoning and Reconsidering Network for Video Captioning. In Pacific Rim International Conference on Artificial Intelligence (pp. 169-182). Springer, Cham.

[122] Yaliniz, G. and Ikizler-Cinbis, N., 2021. Using independently recurrent networks for reinforcement learning based unsupervised video summarization. Multimedia Tools and Applications, 80(12), pp.17827-17847.

[123] Hua, X., Wang, X., Rui, T., Shao, F. and Wang, D., 2022. Adversarial Reinforcement Learning With Object-Scene Relational Graph for Video Captioning. IEEE Transactions on Image Processing, 31, pp.2004-2016.

[124] Kumar, A. and Verma, S., 2021. CapGen: A neural image caption generator with speech synthesis. In Data Analytics and Management (pp. 605-616). Springer, Singapore.

Table 4: Bench mark results of DL models and datasets

Year	Dataset	Models	METEOR	CIDEr	ROUGE	BLEU@4	Core Idea
2020	MSVD	"ResNeXt-101 + ECN (Efficient Convolutional Network)" (Chen et al.2020) [63]	42.1	–	79.7	66.5	Video Captioning
		Inception-Resnet-v2 + C3D (Hou et al. 2020)[70]	34.7	80.1	71.5	47.9	Video Captioning
		InceptionV4 + LSTM-based Guidance Module (Zhang et al.2020) [117]	33.5	83.1	70.7	52.1	Video Summarization
		"2D-CNN + 3D-CNN + compositional decoder" (Perez-Martin et sl.2020) [66]	39.2	107.7	78.3	62.3	Video Description
		"2D-CNN + 3D-CNN + compositional LSTM" (Perez-Martin et al.2021) [67]	41.9	111.5	79.5	64.4	Dense video captioning
	MSR-VTT	"InceptionV3 + C3D and VGG + LSTM" (Xiao et al. 2020) [68]	28.7	48.9	62.3	44.7	Video Captioning
		I3D + Transformer (Liu et al. 2020) [69]	28.5	–	–	41.7	Video Recognition
		"Inception-Resnet-v2 + C3D" (Hou et al.2020) [70]	27.9	45.3	60.1	40.4	Video Captioning
		"2D-CNN + 3D-CNN + compositional decoder" (Perez-Martin et al.2021) [66]	31.4	50.6	64.3	45.5	Video Description
		"2D-CNN + 3D-CNN + compositional LSTM" (Perez-Martin et al.2021) [67]	30.1	48.0	63.1	45.6	Dense video captioning
	Activity Net Captions	Mask RCNN + GRU (Iashin et al.2020) [71]	11.72	–	–	2.86	Dense video captioning
		RestNet-200 + BN Inception + Transformer (Lei et al.2020) [72]	15.57	22.16	5.44	9.78	Video Paragraph captioning
		multimodal neuro-symbolic representations using dictionary learning (Sur et al.2020) [73]	9.78	29.68	20.42	4.01	Video Summarization
		I3D + Transformer (lashin et al.2020) [74]	8.44	–	–	1.88	Dense video

							captioning
		ResNet-200 + LSTM (Suin et al. 2020)[75]	5.7	11.68	–	1.1	Dense video captioning
2020	MPII-MD	GRU + LSTM [Hao et al. 2020) [76]	6.1	10.1	14.6	0.6	Video Recognition
	YouCook2	LSTM (Wang et al. 2020) [93]	13.65	–	–	–	Video Description
		Self-aware composition Transformer + Transformer (Sur et al.2020) [73]	7.34	–	–	0.48	Action Recognition
		multimodal neuro-symbolic representations using dictionary learning (Sur et al.2020) [73]	10.19	50.22	28.17	4.50	Video Summarization
2019	TACOS-MULTILEVEL	JEDDi-Net (Xu et al 2016) [78]	23.9	104.0	50.9	18.1	Video Paragraph captioning
	Activity Net Captions	C3D + LSTM (Zhang et al. 2019) [79]	10.33	12.93	21.21	2.09	Video Summarization
		"ResNet-152 + ResNext-101 (R3D) + (LSTM, GAN)" (Park et al. 2019) [80]	16.48	20.60	–	9.91	Dense video captioning
		"3D CNN + I3D + Bi-directional LSTM + attention-based LSTM" (Wang et al. 2019) [81]	9.96	28.23	21.17	3.68	Video Captioning
2019	MSVD	AGHA [Zhang et al.2019a) [82]	35.3	83.3	–	55.1	Video Captioning
		STAT (Yan et al.2019) [83]	33.5	73.8	–	52.0	Video Summarization
		GoogleNet + VGG + Faster RCNN + LSTM (Guo et al .2019) [84]	35.3	83.3	–	55.1	Video Description
		GoogleNet + Faster RCNN + C3D + LSTM (Yan et al.2019)	33.3	73.8	–	52.0	Dense video captioning
		LSTM + LSTM (Aafaq et al.2019a) [85]	34.3	75.9	–	52.7	Video Captioning
	MSR-VTT	STAT (Chen et al. 2019b) [86]	27.1	44.0	–	39.3	Dense Video Interpretation
		"2D CNN + 3D CNN + Neuron-wise Short Fourier Transform +fully-connected layer + multi-layer GRU" (Chen et al. 2019) [87]	28.4	48.1	60.7	38.3	Video Recognition
		"CNN + temporal deformable convolutional encoder + convolutional decoder+ temporal attention mechanism" (Wang et al.2018) [88]	39.5	42.8	–	38.3	Video Analysis
		"GoogleNet + Inception-Resnet-V2 + C3D + LSTM" (et. Al. Wang 2018) [88]	27.6	47.5	–	42.4	Video Understanding
		"3D CNN + I3D + Bi-directional LSTM + attention-based LSTM" (Wu et al. 2018) [89]	29.4	48.9	62.0	42.2	Dense Video Captioning
2018	MSVD	RecNetlocal (Yang et al.2018) [93]	34.1	80.3	69.8	52.3	Video Understanding
		ResNet + LSTM + LSTM (Wang et al. 2018c) [88]	34.0	74.9	–	51.7	Dense Video Interpretation

		"Joint LSTMs with adversarial learning" (Xu et al.2018) [94]	30.5	–	–	42.9	Video Recognition
		"2D CNN + 3D CNN + LSTM" (Wang et al. 2018a) [88]	33.31	–	–	52.82	Video Analysis
		"Attentive Multi-Grained Encoder (LSTM) + Dual-Stream Decoder (LSTM)" (Yang et al.2018) [93]	34.7	79.4	65.9	53.0	Dense Video Interpretation
	MSR-VTT	RecNetlocal (Wang et al.2020a) [88]	26.6	42.7	59.3	39.1	Dense video description
		Joint LSTMs with adversarial learning (Xu et al.2018) [78]	26.1	–	–	36.0	Video Summarization
		"2D CNN + 3D CNN + LSTM" (Long et al. 2018) [95]	26.58	–	–	38.13	Video Analysis
		"Attentive Multi-Grained Encoder (LSTM) + Dual-Stream Decoder (LSTM)" (Wang et al.2018b) [88]	29.4	46.1	62.3	42.3	Video Interpretation
		ResNet-152 + C3D + LSTM (Wu et al. 2018) [89]	26.7	–	–	39.1	Video Summarization
2018	M-VAD	"Joint LSTMs with adversarial learning" (Yang et al.2018) [93]	6.3	–	–	–	Video Captioning
	MPII-MD	"Joint LSTMs with adversarial learning" (Yang et al.2018) [93]	7.2	–	–	–	Video Captioning
		"Attentive Multi-Grained Encoder (LSTM) + Dual-Stream Decoder (LSTM)" (Xu et al. 2018) [94]	7.9	–	–	1.9	Video Captioning