I.J. Image, Graphics and Signal Processing, **2022, 4, 58-69**
Published Online August 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2022.04.05

Methodology for Translation of Video Content Activates into Text Description: Three Object Activities Action

Ramesh M. Kagalkar
Professor and Dean R&D, Department of CSE, KLE College of Engineering and Technology, Chikodi, Dist. Belagavi, Karnataka, India
Email: rameshvtu11@gmail.com

Received: 11 September 2021; Accepted: 19 March 2022; Published: 08 August 2022

Abstract: This paper presents a natural language text description from video content activities. Here it analyzes the content of any video to identify the number of objects in that video content, what actions and activities are going on has to track and match the action model then based on that generate the grammatical correct text description in English is discussed. It uses two approaches, training, and testing. In the training, we need to maintain a database i.e. subject-verb and object are assigned to extract features of images, and the second approach called testing will automatically generate text descriptions from video content. The implemented system will translate complex video contents into text descriptions and by the duration of a one-minute video with three different object considerations. For this evaluation, a standard DB of YouTube is considered where 250 samples from 50 different domains. The overall system gives an accuracy of 93%.

Index Terms: SVM classification, Computer vision, Gaussian filtering technique, SIFT features.

1. Introduction

Now days a natural language processing (NLP) and computer vision (CV) were located to be growing field of studies and also visible splendid advances of their person objectives of investigating and developing text, and of understanding images and videos at the same time as each fields share a similar association of strategies in artificial intelligence. Late years, in any case, have visible an improvement of enthusiasm for troubles that require a mixture of phonetic and visual records. Ex., Deciphering a photograph with regards to a paper article, adhering to directions related to a graph or a guide, getting slides while tuning in to a lecture. In addition to this, the web gives an immense measure of information that joins etymological and visual data: labeled photos, outlines in paper articles, recordings with captions, and multimodal benefits from online networking. Based on a related work survey we have identified the existing some of the techniques and issues. The state of art of literature survey has been discussed in details.

In [1] gives symmetry functions to become aware of candidate facet additives. Consider that both worldwide and nearby capabilities are extracted from the usage of colorations from unique channels. The extracted capabilities are fed to a logistic classifier for categorization. The overall performance of the machine is evaluated via several text detection and reputation experiments. The [2] present MSR-VTT (status for "MSR Video to Text") which is a brand new big-scale video benchmark for video understanding, especially the rising task of translating video to textual content. Each clip is annotated with approximately 20 natural sentences by way of 1,327 AMT employees. A distinct evaluation of MSR-VTT is offered in evaluation to a complete set of existing datasets, together with a summarization of various cutting-edge video-to-text procedures. In this paper [3] endorse to translate videos immediately to sentences the use of a unified deep neural network with each convolution and recurrent structure. Described video datasets are scarce, and maximum existing techniques have been carried out to toy domains with a small vocabulary of possible phrases. By shifting expertise from 1.2M+ photos with class labels, this technique is capable of creating sentence descriptions of open-area motion pictures with big vocabularies. In paper [4] present an attempt to build a massive-scale JND-primarily based coded video pleasant dataset. This painting describes the subjective test technique, detection and removal of outlying measured records, and the residences of accumulated JND statistics. Finally, the importance and implications of the video set to destiny video coding research and standardization efforts are pointed out.

In [5] uses a natural language description of frames/images. They have used techniques like content planning and surface realization also used is vision detection and classification. Text-statistics from parsing lots of descriptive text and model CRF to predict best image labeling. Generation algorithm is used to compose natural language. [6] Author

has discussed in this paper that by the use of holistic data driven technique, they have generated a text description from videos. It naturally mined from the web scale text corpora and improves the trio choice calculation by giving relevant data and upgrades to a fourfold expansion in movement. [7-12] Tended to acknowledgement of natural human exercises in differing as well as practical video settings. Their first commitment is to deliver comment limitation and to analyze the utilization of film contents for human activities explanation in recordings. They assess elective plans for movement recuperation from contents and appearance central purposes of a substance based classifier. Creators present a novel procedure for a video game plan that develops and broadens a couple.

The author [13] proposed a graph spectral approach for 2D grid images and can be interpret the image as a signal on a graph. He used to apply GSP tools for analysis and processing of the signal in graph spectral domain. This paper [14] address the video summarization technique and it provide in-depth assessment of this pipeline using two popular benchmark data base. It is found that the randomly generated summaries achieve better performance. According this paper [15] author build to overcome an existing work in the area by proposing a novel framework for training the generator against an ensemble of discriminator networks, it can be seen as a one-student/multiple-teachers setting. They have formalized this problem within the full-information adversarial bandit framework and evaluate the capability of an algorithm to select mixtures of discriminators for providing the generator with feedback during learning process. The implemented results shows that it effectively learning a curriculum and also support the claim that weaker discriminators have higher entropy improving modes coverage.

In this paper [16] author identify the challenging to utilize raw diagnostic video data and it takes a long time to process, annotate. So they have come up with a novel, fully automatic video summarization technique to the needs of medical video data. This proposed approach is framed as reinforcement learning problem and produces agents focusing on the preservation of important diagnostic information. Finally this paper concluded that proposed method is superior than the video summarization methods exists one. This paper [17] addresses the reviews the applications of deep learning techniques neuro imaging-based brain disorder analysis. They have highlighted a comprehensive overview of deep learning techniques and also popular network architectures. Then performed deep learning methods for computer-aided analysis of four typical brain disorders, including Alzheimer's disease, Parkinson's disease, Autism spectrum disorder, and Schizophrenia, where the Alzheimer's and Parkinson's disease are neurodegenerative disorders and the Autism spectrum disorder, and schizophrenia, is neuro developmental and psychiatric disorders. And also identify the limitations of existing methods.

This paper [18] addresses and presents some pioneering deep learning models to fuse these multimodal big data. With the increasing exploration of the multimodal big data, there are still some challenges to be addressed. They have highlighted the deep learning for multimodal data fusion to provide readers, regardless of their original community. They have discussed that are widely used are summarized as fundamental to the understanding of multimodal deep learning. Challenges and future scope on this research work also described. In paper [19], we provide a complete comment on current deep education strategies for NER. The proceeding NER resources, inclusive of tagged NER corpora yet off-the-shelf NER tools. Then, we systematically categorize present manufactory primarily based about taxonomy along 3 axes: allotted representations because of input, adherence encoder, yet tag decoder. Next, we are metering the most consultant methods because of current applied techniques concerning awful learning between recent NER hassle settings yet applications. Finally, we present readers along the challenges confronted by NER systems and outline future instructions within this area.

In paper [20] writer cope with literature survey on deep gaining knowledge of and artificial intelligence strategies for autonomous using, additionally discussed and supplied AI-based self-driving architectures, convolutional and recurrent neural networks, in addition to the deep reinforcement studying paradigm. They have identified and look at each the modular notion-making plans-movement pipeline the use of deep mastering strategies and End2End systems, which directly map sensory facts to guidance commands. Discussed the cutting-edge challenges, electricity and barriers, in this work. In this paper [21] authors cope with the Video summarization techniques. Right here it picks out the important thing frames to symbolize the effective contents of a video collection. Several strategies available are extracting the static pictures of films, where dynamic motion video content material is remain untouched. This paper copes with the unique framework for an efficient video content as well as video motion summarization. It constitutes a terrific motion summarization result.

This paper [22] evaluates five publicly-to be had summarization algorithms below a big-scale experimental putting with 50 randomly-created databases. The outcomes said in the papers are not continually congruent with their overall performance on the massive-scale test. This paper [23] presents a web carrier that supports the automated era of video summaries for person-submitted films. The advanced internet application decomposes the video into segments, evaluates the health of every segment to be covered in the video summary and selects appropriate segments till a pre-described time finances is stuffed. In this paper [24] writer deal with a technique which takes a text-primarily based question as input and generates a video précis corresponding to it. It makes use of modeling video summarization as a supervised getting to know hassle and advocate an cease-to-stop deep learning primarily based technique for question-controllable video summarization to generate a question-structured video précis. The proposed approach consists of a video précis controller, video précis generator, and video summary output module. This paper additionally mentioned a

dataset that incorporates body-primarily based relevance score labels. Based totally on experimental result, the textual content-based totally question enables manage the video summary. It additionally improves their performances.

In this paper [25] given a detailed examine of sampling indicators on graphs, with the goal of constructing an analog of sampling for widespread signals within the time and spatial domain names, has attracted big interest recently. Beyond adding to the growing theory on graph signal processing (GSP), sampling on graphs has various promising packages. On this paper creator tried to discussed idea and capacity applications as well because the current development on sampling over graphs. This paper [26] endorse three novel curriculum gaining knowledge of strategies for schooling GANs. All techniques are first based on rating the schooling pics via their difficulty scores that are estimated by means of a trendy photo trouble predictor. The primary method is to divide photos into step by step extra difficult batches. Whereas 2nd method introduces a singular curriculum loss function for the discriminator that takes into consideration the difficulty scores of the real images. The third strategy is based on sampling from an evolving distribution, which favors the simpler pics at some point of the preliminary education levels and regularly converges to a uniform distribution. The experiments indicate that all techniques offer quicker convergence and proper outcomes.

This paper [27] suggest a new algorithm this is composed of segmentation, background initialization, graph production, unseen sampling, and a semi-supervised mastering technique inspired through the concept of healing of graph indicators. Their algorithm has the gain of requiring less classified statistics than deep getting to know techniques while having aggressive consequences on each static and shifting digicam films. Therefore this set of rules is also tailored for Video item Segmentation (VOS) responsibilities and is evaluated on six publicly to be had datasets outperforming and challenging situations. In this paper [28] author analyzed and reviewed using deep mastering algorithms for Cyber safety packages. They have studied more than 80 research paper at the equal domains are referred and analyzed, wherein recognized every unique technique and mentioned with its algorithms, platforms, dataset, and potential benefits. From the experimental evaluation, they've noticed that the deep learning model progressed the accuracy, scalability, reliability, and overall performance [29-33].

Motivation of this work like analysis of video information where no of objectives activities and its action identification, analysis and recognization is not a easy task and i studied the many relevant research papers work it is found that remain untouched, so we have come up with implementation work for achieving few practical results.

We made the following contributions:

- Here it analyzes the content of any video to identify the number of objects in that video content, what actions and activities are going on has to track and match the action model then based on that generate the grammatical correct text description in English is discussed.
- It uses two approaches, training, and testing. In the training, we need to maintain a database i.e. subject-verb and object are assigned to extract features of images, and the second approach called testing will automatically generate text descriptions from video content.
- The implemented system will translate complex video contents into text descriptions and by the duration of a one-minute video with three different object considerations.
- For this evaluation, a standard DB of YouTube is considered where 250 samples from 50 different domain.
- The overall system gives an accuracy of 93%.

Subsequently, the summary of the paper is organized as sections, in section II, we discussed the Methods and materials used in this paper. Where in section III had given detailed description database consideration. In section IV result analysis and discussion. While section V provides the conclusion and prospective work.

2. Methodology

The frame work presents a procedure for producing characteristic language represetation of long length recordings by distinguishing the item and activities for depicting recordings. Here framework comprises of two significant modules training and testing.

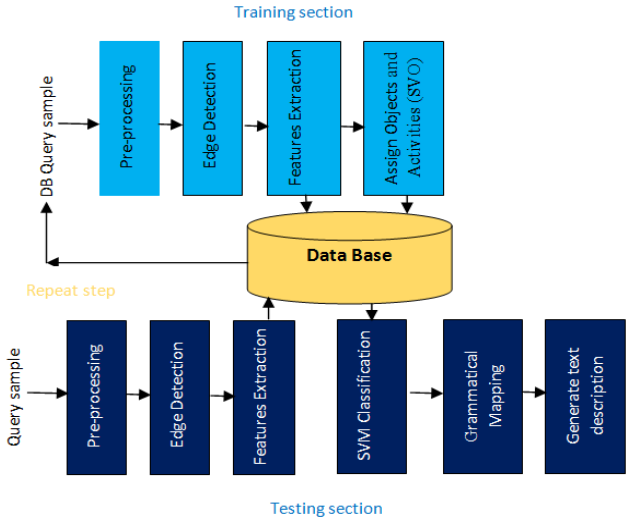

Fig. 1. Overview of system.

2.1 Training section

In this section we have considered the set of training video samples of 3 objects considerations, where from each sample we have extracted certain features, objects and its activity description and are stored in a database such that entire training video samples features are stored one by one. The fig 1 shows the proposed overview of system. The training process undergoes step by step operation such as selection of training samples videos, and then it is split into images/frames. The pre-processing is applied on to video samples to check video content quality and if any blur video should be eliminated and make it clear identification of edge of object by applying Gaussian filtering technique then edge detection is applied to calculate the edge of object. Finding the number of objects in each frame and considering only dissimilar frames in the video. Finally from the frame which underwent edge detection, where calculating 10 features (SIFT features) like shape, color, texture, contrast, object position and homogeneity. Then assign subject verb and object information to these features values and also coordinate values are stored in separate files in the database, like that all samples from the database are to be trained.

2.2 Testing section

In testing considered a video sample from a trained database then it also undergoes a similar process as training such as pre-processing, edge detection using canny edge detector, extracted features like shape, color, texture, contrast, object position and homogeneity. All these features values are stored in a database, then we use a classifier to identify the meaning of the frame then based on features values text generation has been performed and also apply grammatical mapping to check the grammar from the generated text.

A major pre-processing step taken was to reduce lighting condition effects by first applying masks to the image, and then we resized the image to 224 ×224. On the resized image we applied Gaussian blur to enhance the image. We cropped out uninformative areas from the image and kept only the necessary parts. The Gaussian blur feature is obtained by blurring (Smoothing) an image using a Gaussian function to reduce the noise level. It can be considered as a non-uniform low-pass filter that preserves low spatial frequency and reduces image noise and negligible details in an image. Canny edge detection is a technique to extract useful structural information from different vision objects and dramatically reduce the amount of data to be processed. It has been widely applied in various computer vision systems. Canny has found that the requirements for the application of edge detection on diverse vision systems are relatively similar. Thus, an edge detection solution to address these requirements can be implemented in a wide range of situations. The general criteria for edge detection include: Detection of edge with low error rate, which means that the detection should accurately catch as many edges shown in the image as possible the edge point detected from the operator should accurately localize on the center of the edge. A given edge in the image should only be marked once, and where possible, image noise should not create false edges.

To satisfy these requirements Canny used the calculus of variations – a technique which finds the function which optimizes a given functional. The optimal function in Canny's detector is described by the sum of four exponential terms, but it can be approximated by the first derivative of a Gaussian. Among the edge detection methods developed so far, Canny edge detection algorithm is one of the most strictly defined methods that provides good and reliable detection. Owing to its optimality to meet with the three criteria for edge detection and the simplicity of process for implementation, it became one of the most popular algorithms for edge detection. SIFT features is a feature detection algorithm in Computer Vision. it helps locate the local features in an image, commonly known as the ‘key points‘ of the image. These key points are scale & rotation invariant that can be used for various computer vision

applications, like image matching, object detection, scene detection, etc. We can also use the key points generated using SIFT as features for the image during model training. The major advantage of SIFT features, over edge features or hog features, is that they are not affected by the size or orientation of the image.

1. Training

Input: Set of video samples.
Output: Feature Extraction with its description of a video samples and stored in database.
Steps:
Start
Step 1- Consider an input video sample and extract frames.
Step 2- Pre-processing is performed to remove noise, blur from the surface of frames by using Gaussian filtering algorithm.

$$G(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2)/2\sigma^2} \quad (1)$$

Hear x is the distance from the origin in the horizontal axis, y is the gap from the origin in vertical axis σ is the same old deviation of the distribution.

Step 3- Edge detection to identify edges by using Canny edge detector.
First it smooths the image i.e. suppress as much noise as possible, without destroying the true edges. Then it finds the gradient by using equation.

$$G|=\sqrt{Gx^2 + Gy^2} \quad (2)$$

$$|G| = |Gx| + |Gy \quad (3)$$

Where ,Gx and Gy are the gradients of x- and y-directions respectively.
Orientation of the edges can be finding by using following equation

$$G(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2)/2\sigma^2} \quad (4)$$

Step 4- Features extracted from images by using SIFT and store into the database with its textual description. It extracts 10 features from video sample;
The commonly used features are:

1. Shape: It is the similarity between shapes and their features.
2. Color: Combination of three colors: Red, Green and Blue.
3. Texture: Used to classify and recognize objects.
4. Contrast (C): It determines the brightness of image.
5. Objects position: These are the domain-specific features.

- Orientation
- Location
- Head position
- Hand position

6. Homogeneity: Used to detect low level image features such as edges and corners. It determines the location of feature points by using following equation.

$$L(x, y, \sigma) = G(x, y, \sigma)* I(x, y) \quad (5)$$

Where,

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

Right here, x is the gap from the starting place within the horizontal axis. Y is the space from the origin in vertical axis, i is the depth of the pixel, σ is the usual deviation of the distribution.

Step 5- Assign the subject, verb and object and store all features in database.
Step 6- Repeat the step 1 for up to 'n' video samples for training.
Stop

2. Testing

Input- Testing video sample.
Output-Text Generation in English language with correct grammar.
Start
Step 1- Consider an input video sample. Extract frames from the video sample.
Step 2- Pre-processing is performed to remove noise, blur from the surface of frames by using Gaussian filtering algorithm.

$$G(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2)/2\sigma^2} \quad (6)$$

Hear x is the distance from the origin in the horizontal axis, y is the gap from the origin in vertical axis σ is the same old deviation of the distribution.
Step 3- Edge Detection to identify edges by using Canny Edge Detector.
First it smooths the image i.e. suppress as much noise as possible, without destroying the true edges. Then it finds the gradient by using equation.

$$|G|\sqrt{Gx^2 + Gy^2} = \quad (7)$$

$$|G| = |Gx| + |Gy| \quad (8)$$

Where Gx and Gy are the gradients of x- and y-directions respectively.
Orientation of the edges can be finding by using following equation.

$$G(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2+y^2)/2\sigma^2} \quad (9)$$

Step 4- Features extracted from images by using SIFT and store into the database with its textual description. It extracts 10 features from video sample; the commonly used features are:

1. Shape: It is the similarity between shapes and their features.
2. Color: Combination Of three colors: Red, Green and Blue.
3. Texture: Used to classify and recognize objects.
4. Contrast (C): It determines the brightness of image.
5. Objects position: These are the domain-specific features.
   - Orientation
   - Location
   - Head position
   - Hand position
6. Homogeneity: Commonly used to detect low level image features such as edges and It determines the location of feature points by using following equation.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y). \quad (10)$$

where

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

Right here, x is the gap from the starting place within the horizontal axis. Y is the space from the origin in vertical axis, i is the depth of the pixel, σ is the usual deviation of the distribution.
Step 5- Classify testing video features with training video features, and apply subject, verb and object.
Step 6- If features are matched then it generates text description of input video sample.
Stop

3. Database Consideration

For the system implementation we have considered approximately 200 videos from diverse domain names. And we also use the standard English portion of the YouTube data collected, consisting of short videos each with multiple natural-language descriptions. This data was ensured that the test data only contained videos in which we can potentially detect objects. To study the textual content extraction process, videos and its description are to be stored into the database to offer satisfactory results. These motion images are divided into various classes like marketplace, water park, playground, hospital, railway station, traffic signal, college area, airport etc. In table 1 dataset gives detailed explanation of each video sample name, domain, number of video samples and description of video samples is elaborated. The training datasets are prepared based on the domain with the number of objects in each video sample. All this video samples are video captured using Q4N Handy Video Recorder, with ultra-flexible video camera designed for acting artists, super a hundred and sixty-degree huge-attitude lens (f2. 0/16. 6 mm) presents an extended subject of view, aid for five HD video modes, up to 2304 x 1296 pixels (3m HD), in addition to 2 modes, video bitrates up to 24 M bit/s for prolonged photo first-class and body quotes of up to 60 f/s.
https://github.com/pedropro/TACO/blob/master/data/all_image_urls.csv

Table 1. Training dataset.

Sr. No.	Domain	Number of video samples	Description of video
1	Railway Station	Railway_Station_Video 1	This video is of railway station. A train is going from the platform and passengers are waiting for the train.
		Railway_Station_Video 2	There are peoples enquiring about the trains.
		Railway_Station_Video 3	Peoples are trying to get into the train.
		Railway_Station_Video 4	Peoples are going from one platform to another.
		Railway_Station_Video 5	A train is arriving on the station.
		Railway_Station_Video 1	This video is of railway station. A train is going from the platform and passengers are waiting for the train.
		Railway_Station_ Video 2	There are peoples enquiring about the trains.
2	College	College_ Video 1	This video is of college. Professor conducting practical.
		College_Video 2	There is meeting in the principle room. Teaching staff are attending meeting.
		College_Video 3	Staff is conducting practical and students.
		College_Video 4	Computer practical is going on in the computer lab.
		College_Video 2	There is meeting in the principle room. Teaching staff are attending meeting.
3	Shopping	Shopping_Video 1	A lady is buying jewellery.
		Shopping_Video 2	There is shop of footwear and girl is buying sandals.
		Shopping_Video 3	A girl is buying clothes and talking to the salesman.
		Shopping_Video 4	Two girls talked with the salesman and looking for an auto.
		Shopping_Video 5	A girl is paying to the seller and talked with her friend.
		Shopping_Video 1	A lady is buying jewellery.
4	Play Ground	Play_ground_Video 1	Children playing with football.
		Play_ground_Video 2	Some children play on a ride.
		Play_ground_Video 3	A boy enjoying swing.
		Play_ground_Video 4	Four boys playing on a ground.
		Play_ground_Video 5	Some children are running.
5	Airport	Airport_Video 1	Girls are running in a race.
		Sports_Video 2	Boys are playing with football.
		Sports_Video 3	There is badminton court and two boys are playing badminton.
		Sports_Video 4	This video is of cricket and players are playing.

4. Experimental results analysis and discussion

In this section, we have selected a few video samples from databases of different domains taken as an input and computed time duration of videos, processing time, expected and actual results are computed as shown in table 2. Also, the comparison of actual output and expected output are discussed further.

Table 2. Result analysis of implemented system.

Sr. No.	Video sample domain	Time duration (Seconds)	Processing time (Milliseconds)	Expected result	Actual result
1	Railway_Station	07	10750	This video is of railway station, a train is arriving on the platform. Passengers are waiting on the platform.	Passengers wait for the train. Train arrives at the platform. People gets into the train.
2	Market	30	34055	A man is selling food and gave teapot to other man and serves tea.	A man sells food on road. A man prepares tea. Man gives teapot.
3	Shopping	27	26799	There is a video of shopping area. There is a girl walking in a market area.	Girl walks on a road. She buys some jewellery. Salesman sells jewellery.
4	Airport	10	13200	This video is of airport. Passengers are waiting for airbus. Some people goes upstairs.	People goes upstairs in the aero plane. Some people wait for the airbus.
5	Water_Park	28	31580	Video is of water park. Water tank is full of water. Water is falling from water tank. Some children enjoying water fall.	Girl plays in water. There is water park. Water falls on rides. Girl is playing on ride. Some children are playing in water.

In table 2, video 1 is of the railway station area. For experiments, we have taken a 7 seconds video. In this video, a train is arriving on the platform and some passengers are getting into the train. People are waiting on the platform. The video contains so many objects like platform; trains, and crowds of people. There are approximately 20 objects in the video but only those objects are considered which describe the scenery of a video and other objects are discarded. This video gives the actual output "Passengers wait for the train. Train arrives at the platform. People get into the train" And the expected output is "This video is of a railway station; a train is arriving on the platform. Passengers are waiting on the platform.". This video 1 takes 10750 milliseconds for processing. Similarly, video 2, video 3 up to video 5 results are shown.

Table 3. Results summary of presented system for three objects.

Sr. No.	Video samples Domain	Duration in Seconds	Number of Objects	Processing Time in Millisecond			
				Gaussian	Canny	SIFT	Recognition Rate in %
01	Market	30	03	473	4482	29100	Expected words = 14 Actual words = 13 Accuracy = 92.8%
02	Shopping	27	03	354	3755	22690	Expected words = 14 Actual words = 12 Accuracy = 85.7%
03	Water park	28	03	1050	4130	26400	Expected words = 22 Actual words = 22 Accuracy = 100%

In table 3, Video samples are taken from a market domain, processing time and recognition rate is shown. For this video expected words are 14 and the system generated words are 13, so the recognition rate calculated is 92.8%. Similarly, for shopping domain 85.7 and for water park is 81%. Recognition fee is calculated based at the end result generated. For video pattern 1 anticipated word are 18 and the words generated by way of the system are 15 so the recognition rate is 88%. Similarly, mall and playground reputation calculated are 92%, 80% and 93. 7%.

4.1 Analysis of Frame Extraction

Here analysis of frame and selection of number of frames are discussed The video samples of various domain names are considered as an input and table 4 shows the whole range of frames extracted for each video. Distinct frames are considered and the similar frames are discarded.

Table 4. Number of frames comparison for object 3.

Sr. No.	Domain	Number of objects	Number of frames extracted
1	Market	3	8
2	Shopping	3	8
3	Playground	3	3

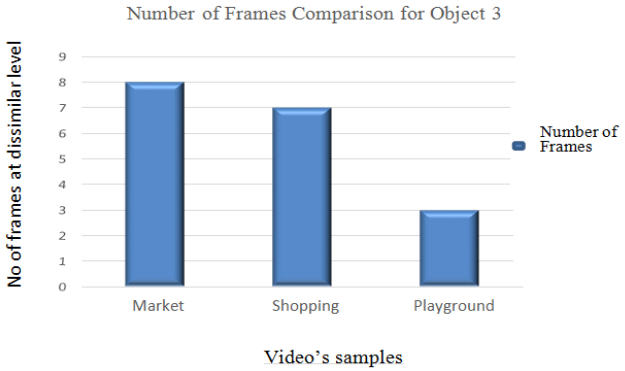

Fig. 2. Shows a comparison of No. of frames Vs videos samples for object 3.

The fig. 2 shows a comparison of No. of frames Vs videos samples for object 3 in graph of No of frames at dissimilar level versus No. of video's samples.

Avg. No. of frames extracted for object 3 = (8+8+3)/63 = 6 frames.

So, the Avg. No. of frames extracted is 6 frames.

4.2 Analysis of Processing Time

It is the average time taken by each algorithm i.e. Gaussian, Canny and SIFT. Processing time is related to the video size and the number of objects present in each frame shown in table 5. The larger the video size, the higher the processing time.

Table 5. Time comparison for object 2.

Sr. No.	Domain	Time in Milliseconds
1	Market	30115
2	Shopping	45000
3	Playground	31180

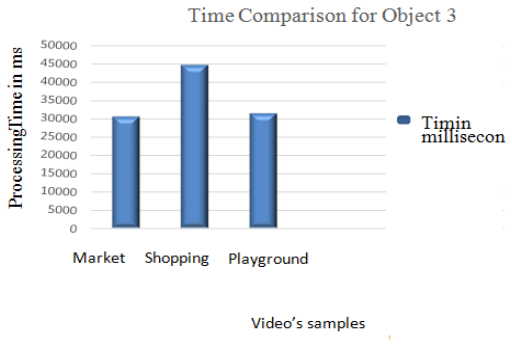

Fig. 3. Graph for time comparison for object 3.

The fig. 3 shows a graph for time comparison for object 3 with the comparison of processing time vs. No. of videos samples for object 3.

Avg. No. of frames extracted for object 3 = (8+8+3)/ 63 = 6 frames

So, the Avg. No. of frames extracted is 6 frames.

4.3 Analysis of Recognition Rate Recognition rate for object 3

The standard device performance may be measured with the assistance of a wide variety of processing time and rate of recognization. Total processing time is the common time taken by means of every algorithm i.e. Gaussian, Canny and SIFT. Processing time is directly associated with the video length and the variety of objects present in each frame is shown in table 6. More the video period, the higher the processing time might be.

Table 6. Recognition rate of object 3.

Sr. No.	Domain	Recognition Rate in %
1	Market	93
2	Shopping	86
3	Playground	100

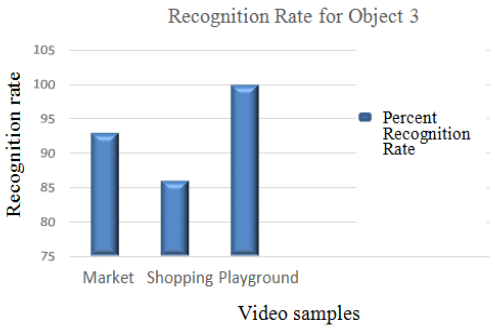

Fig. 4. Graph for recognization rate of object 3.

The fig. 4 shows a graph for recognization rate of object 3 with the comparison of recognization rate vs No. of videos samples for object 3.

Average recognition rate for object 3 = (93+ 86+100)/3 = 93%

4.4 Comparison of existing techniques

The proposed work is going to compare the existing methods for video content analysis results obtained. Table 7 shows the methods with its performance accuracy. It evaluate the three essential phases. The first-rate strategies phase, where pioneering visible account lookup busy classic CV or NLP strategies in conformity with forward notice entities (objects, actions, scenes) into videos or afterwards fit them in conformity with par condemnation templates. The statistical methods phase, as attached statistical techniques after deal including exceedingly larger datasets. This phase lasted for a relatively short time. Finally, the deep learning phase, which is the current state of the art and is believed to have the potential to solve the open domain automatic video description problem with detailed survey of the methods in each category.

Table 7. Comparison of existing techniques.

Sr. No.	Methods	Performance analysis
1	Classical methods: Uses classical CV and NLP methods to first detect entities (objects, actions, scenes) in videos and then fit them to standard sentence templates	92% accuracy with standard data base.
2	The statistical methods phase, which employed statistical methods to deal with relatively larger datasets. This phase lasted for a relatively short time	86% accuracy with standard local data base consideration.
3	The deep learning phase, which is the current state of the art and is believed to have the potential to solve the open domain automatic video description problem	90% accuracy with real time video data consideration.
4.	Uses two approaches, training, and testing then performed Use Gaussian filtering technique, Canny edge detection is a technique, SIFT Features, Classify testing video features with training video features, and apply subject, verb and object.	93% accuracy with local data base.

5. Conclusion

The implemented system performs object detection, action reorganization and feature extraction from the given video sample. It performs frame extraction, pre-processing, edge detection and features extraction and extracted features are stored in database. Each video splits into frames at one-second intervals and the filtering, shape detection techniques are applied on every frame. Features are extracted using SIFT algorithm and these features are used for comparison of testing with the training video and gives the description of video. The performance of the implemented system can be calculated and gives the accuracy of 93%.

References

[1] Longfei Qin, Palaiahnakote Shivakumara, Tong Lu, Umapada Pal and Chew Lim Tan, "Video Scene Text Frames Categorization for Text Detection and Recognition", 23rd International Conference on Pattern Recognition (ICPR), México, December 4-8, 2016.

[2] Jun Xu,Tao Mei, Ting Yao and Yong Rui, "MSR-VTT:ALargeVideoDescriptionDatasetforBridgingVideoandLanguage", IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[3] Subhashini Venugopalan Marcus Rohrbach ,"Translating Videos to Natural Language Using Deep Recurrent Neural Networks" Annual Conference of the North American Chapter of the ACL, pages 1494–1504, June 2015.

[4] Haiqiang Wang, LoannisKatsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, "Videoset: A Large –Scale Video Quality Dataset Based on JND Measurement", Elsevier 2017.

[5] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, Siming Li, Y. Choi, A. C. Berg, and Tamara L. Berg, "Baby Talk: Understanding and Generating Simple Image Descriptions", IEEE Transaction on pattern analysis and machine intelligence, vol. 35, no. 12, Dec 2013.

[6] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, and S. Guadarrama, "Generating Natural-Language Video Descriptions Using Text-Mined Knowledge", 2013

[7] Laptev, I., and Perez, P., "Retrieving Actions in Movies", In Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV), pp. 1–8, 2007.

[8] Laptev, I. Marszalek, M. Schmid, C. and Rozenfeld, B., "Learning Realistic Human Actions from Movies", In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2008.

[9] Edke, Vandana D., and Ramesh M. Kagalkar. "Review Paper on Video Content Analysis into Text Description." International Journal of Computer Applications National Conference on Advances in Computing. 2015.

[10] Kagalkar R.M., Khot P., Bhaumik R., Potdar S., Maruf D. (2020) SVM Based Approach to Text Description from Video Sceneries. In: Jyothi S., Mamatha D., Satapathy S., Raju K., Favorskaya M. (eds) Advances in Computational and Bio-Engineering. CBE 2019. Learning and Analytics in Intelligent Systems, vol 15. Springer, Cham. https://doi.org/10.1007/978-3-030-46939-9_52.

[11] D. Edke, Vandana, M. Kagalkar, Ramesh Video Object Description of Short Videos in Hindi Text Language, International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 12, Number 2 (2016), pp. 103-116 © Research India Publications.

[12] Wankhede, V., Kagalkar, R.M. Conference Paper, "Efficient approach for complex video description into English text", Proceedings of 2017 International Conference on Intelligent Computing and Control, I2C2 2017, 2018, 2018-January, pp. 1–7.

[13] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph Spectral Image Processing," Proc. of the IEEE, vol. 106, no. 5, pp. 907–930, 2018.

[14] M. Otani, Y. Nakahima, E. Rahtu, and J. Heikkil¨a, "Rethinking the Evaluation of Video Summaries," in 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.

[15] T. Doan, J. Monteiro, I. Albuquerque, B. Mazoure, A. Durand,J. Pineau, and D. Hjelm, "On-line Adaptative Curriculum Learning for GANs," in Proc. of 2019, AAAI Conf. on Artificial Intelligence, March , 2019.

[16] T. Liu, Q. Meng, A. Vlontzos, J. Tan, D. Rueckert, and B. Kainz, "Ultrasound video summarization using deep reinforcement learning," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer International Publishing, pp. 483–492, 2020.

[17] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," Frontiers in Neuroscience, vol. 14, pp. 779, 2020. https://www.frontiersin.org/article/10.3389/fnins.2020.00779.

[18] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," Neural Computation, vol. 32, no. 5, pp. 829–864, 2020.

[19] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE Trans. on Knowledge and Data Engineering, pp. 1–1, 2020.

[20] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21918.

[21] C. Huang and H. Wang, "A Novel Key-Frames Selection Framework for Comprehensive Video Summarization," IEEE Trans. on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 577–589, 2020.

[22] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, “Performance over Random: A Robust Evaluation Protocol for Video Summarization Methods,” in Proc. of the 28th ACM Int. Conf. on Multimedia (MM ’20). New York, NY, USA: ACM, pp 1056–1064, 2020.
[23] C. Collyda, K. Apostolidis, E. Apostolidis, E. Adamantidou, A. I. Metsai, and V. Mezaris, “A Web Service for Video Summarization,” in ACM Int. Conf. on Interactive Media Experiences (IMX ’20). New York, NY, USA: ACM, pp 148–153, 2020.
[24] J.-H. Huang and M. Worring, “Query-controllable video summarization,” in Proc. of the 2020 Int. Conf. on Multimedia Retrieval, ser. ICMR ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 242–250. [Online]. Available: https://doi.org/10.1145/3372278.3390695
[25] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, “Sampling Signals on Graphs: From Theory to Applications,” IEEE Signal Processing Magazine, vol. 37, no. 6, pp. 14–30, 2020.
[26] P. Soviany, C. Ardei, R. T. Ionescu, and M. Leordeanu, “Image Difficulty Curriculum for Generative Adversarial Networks (CuGAN),” in 2020 IEEE Winter Conf. on Applications of Computer Vision (WACV), 2020, pp. 3452–3461.
[27] J. H. Giraldo, S. Javed, and T. Bouwmans, “Graph Moving Object Segmentation, ”IEEE Trans. on Pattern Analysis and Machine Intelligence,pp. 1–1, 2020.
[28] P. Dixit and S. Silakari, “Deep learning algorithms for cyber security applications: A technological and status review,” Computer Science Review, vol. 39, p. 100317, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013720304172.
[29] Mesut. Polatgil, "Investigation of the Effect of Normalization Methods on ANFIS Success: Forestfire and Diabets Datasets", International Journal of Information Technology and Computer Science(IJITCS), Vol.14, No.1, pp.1-8, 2021. DOI: 10.5815/ijitcs.2022.01.01.
[30] Prashengit Dhar, Md. Shohelur Rahman, Zainal Abedin, "Classification of Leaf Disease Using Global and Local Features", International Journal of Information Technology and Computer Science(IJITCS), Vol.14, No.1, pp.43-57, 2022. DOI: 10.5815/ijitcs.2022.01.05.
[31] Oksana Babich, Viktor Vyshnyvskiy, Vadym Mukhin, Irina Zamaruyeva, Michail Sheleg, Yaroslav Kornaga, "The Technique of Key Text Characteristics Analysis for Mass Media Text Nature Assessment", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.1, pp. 1-16, 2022.DOI: 10.5815/ijmecs.2022.01.01.
[32] Monika Arora, Indira Bhardwaj, "Artificial Intelligence in Collaborative Information System", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.1, pp. 44-55, 2022.DOI: 10.5815/ijmecs.2022.01.04.
[33] Stephen Akuma, "Eye Gaze Relevance Feedback Indicators for Information Retrieval", International Journal of Intelligent Systems and Applications(IJISA), Vol.14, No.1, pp.57-65, 2022. DOI: 10.5815/ijisa.2022.01.05.

Author’s Profile

Dr. Ramesh M. Kagalkar is an academician with 19 years of experience, 30+ International publications, 15+ International conference paper presented, 6+ published patents (Two are under process of granting stage), Received research grant of total Rs. 8 Lakh from TEQIP Competitive Research Grant from VTU, Belagavi, Author of 3 academic Text book to his credit. His paper citations are like Citation: 232, h-index: 08 and i-index: 08. He earned a Bachelor of Engineering (CSE-2001) Gulbarga University, Gulbarga, Karnataka, Master of Technology (M.Tech-2006) from Visvesvaraya Technological University, Belgaum, Karnataka and a Doctorate (Ph.D.) in Computer and Information Science (Ph.D-CISc-2019) from Visvesvaraya Technological University, Belgaum, Karnataka. He has guide 20+ PG projects and 30+ UG projects. Ability to handle innovative major project of UG final year students and also define research topic for guiding the Ph.D students. He is providing research solution to different domains such as Image, Video, Audio, etc. He is presently working on social problem to provide technical solutions for Blind, Deaf, and Disable, Aged individual, women and children safety service system projects. His research interest is in the areas of Image, Video and Audio processing.

How to cite this paper: Ramesh M. Kagalkar, "Methodology for Translation of Video Content Activates into Text Description: Three Object Activities Action", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.14, No.4, pp. 58-69, 2022. DOI:10.5815/ijigsp.2022.04.05

