

# VIDEO OBJECTS DESCRIPTION IN HINDI TEXT LANGUAGE

Vandana D. Edke<sup>1</sup> and Ramesh M. Kagalkar<sup>2</sup>

<sup>1</sup>2nd year M.E Student of Computer Eng. Dept,  
Dr. D Y Patil School of Eng. and Technology, Pune

<sup>2</sup>Research Scholar and Asst. Professor, Computer Eng. Dept,  
Dr. D Y Patil School of Eng. and Technology, Pune

## ABSTRACT

*Video activity recognition has grown to be a dynamic location of analysis in latest years. A widespread information-driven approach is denoted in this paper that produces descriptions of video content into textual content description inside the Hindi language. This method combines the final results of modern item with "real-international" records to pick the in all subject-verb-object triplet for depicting a video. The usage of this triplet desire technique, a video is tagged via the trainer, mainly, Subject, Verb, and object (SVO) and then this data is mined to improve the result of checking out video clarification by using pastime as well as item identity. Contrasting preceding approaches, this method can annotate arbitrary videos deprived of wanting the large series and annotation of a similar schooling video corpus. The proposed work affords initial and primary text description within the Hindi language that is producing easy words and sentence formation. But the fundamental challenging attempt on this work is to extract grammatically accurate and expressive text records in Hindi textual content regarding video content.*

## KEYWORDS

*Natural-language processing, Surface realization stage, Stanford dependency parser, Video processing*

## 1. INTRODUCTION

Increasing sharing of public image and video on websites, together with "Flickr" and "YouTube" gives you a large corpus of unstructured image and video statistics over the internet. searching and retrieving visible proof from the net, but, has been frequently limited to the use of meta-information, person- annotated tags, subtitles and surrounding textual content (e.g. the image seek engine used by Google [1]). Merging natural-language processing (NLP) with computer imaginative and prescient to generate Hindi descriptions of visible information is a good sized place of active research.

The paper offers a singular technique to producing a simple sentence for describing a brief video that: First, identifies the maximum expected subject, verb and object (SVO) with the aid of a method for a mix of visible object and activity detectors and text-mined records to evaluate the chance of SVO triplets. From a natural-language technology (NLG) angle, this is the content material arrangement stage. Secondly, assumed the selected SVO triplet, it makes use of a easy template based totally technique to generate candidate sentences that are then ranked with the aid of a statistical language version prepared on web-scale statistics to get the exceptional worldwide portrayal. The proposed technique can be watched as a holistic data-driven three-step process where firstly of all observe the objects and activities the use of cutting-edge visible popularity algorithms. Except, consolidate these regularly noisy detections with an assessment of real likelihood, that's obtain via mining SVO triplets from widespread scale net corpora. In the end,

those triplets are used to create candidate sentences that are then ranked for plausibility as well as grammaticality. The subsequent natural-language descriptions can be usefully employed in claims inclusive of semantic video seek and summarization and offering video interpretations for the outwardly impeded. Using vision models alone to estimate the best subject and object for a given interest is difficult, especially while handling inspiring real-world YouTube motion pictures because it calls for an extensive commented on video corpus of comparative SVO triplets. The supplied method is complicated in annotating arbitrary short movies using off-the-shelf visible detectors, disadvantaged of the engineering effort vital to constructing domain-precise hobby models. The main contribution is inclusive of the pragmatics of numerous entities' possibility of being the problem/object of a given interest, discovered from net-scale textual content corpora. For example, animate gadgets like humans and animals are much more likely to be topics likened to inorganic gadgets like balls or television monitors. Similarly, certain items are more likely to function as topics or objects of specific sports, e.g., "using a horse" verses "riding a house Picking the great verb may also likewise key perceiving activities for which no unequivocal schooling facts has been conveyed. For example, reflect on consideration on a video with a person strolling his pet dog. The item locators may additionally apprehend the person and the canine; but the activity indicators may basically have the greater general motion, "pass" in their training information. In such situations, actual world pragmatics could be very useful in signifying that "stroll" is fine used to designate a person "moving" along with his canine. This process is referring as verb enlargement. Eventually, effects are computed by the usage of real global brief duration movies. On this paper, a description of video content material into textual content the usage of Hindi language is proposed.

## 2. LITERATURE SURVEY

Within the literature assessment, we're going to debate topical strategies over the video text reputation: underneath in literature we're debating a number of them.

V. Edke and R. Kagalkar [3] defined evaluate on video content material observe into textual content description. Accordingly this paper presented three important contributions to activity recognition from video. First of all, they added a single mechanism for robotically discovering films hobby classes from herbal-language descriptions. Secondly, a current hobby popularity scheme is progressed abuse object context along with relationships among items and activities. Ultimately, indicates language manner is acquainted routinely extracting the needful records approximately the relationship among items and sports from a corpus of popular text.

R. Hiremath and R. Kagalkar [4] presented assessment on sign language recognition for the important thing finding of the comparative analysis of similar techniques and additionally for era utilized in imaginative and prescient primarily based hand gesture recognition.

Chang, C. et al. [6] supplied complete implementation details of support Vector Machines called LIBSVM. Though, this article does not a goal to give an explanation for the practical use of LIBSVM for recommendations of using LIBSVM. De Marneffe [7] depicts a framework for eliminating typed reliance parses of English sentences on account that expression structure parses. That allows you to capture basic family members taking place in corpus texts that may be hazardous in real-world applications, numerous Noun phrase (NP) family members are covered within the set of grammatical family members used.

Dings, D. et al. [8] assessment previous observe on audio as well as video processing, and describe the subject-oriented Multimedia Summarization (TOMS) challenge the usage of natural Language technology: given a hard and fast of mechanically mined features from a video. A topic-oriented multimedia summarization (TOMS) framework will consequently create a passage

of commonplace dialect, which outlines the vital facts in a video having an area with a particular factor range, and offers elucidations to why a video was coordinated, recovered, and so forth. They display this as a primary stage toward schemes with a purpose to be capable of discriminate visually similar, however semantically exclusive motion pictures, partner two motion pictures and deliver composed yield or summarize a widespread number of films without a moment's put off. Authors present method of determining the TOMS trouble. They dispose of visible concept components and ASR interpretation and enhance a Template-based totally herbal Language technology (NLG) Scheme to create a composed relating in view of the mined factors. Authors moreover suggest conceivable designs plans for constantly assessing and refining TOMS frameworks, and present consequences of a pilot design of preliminary framework [8]. Ali Farhadi [9] et al. described a machine that may compute a rating related to an image to a sentence. This score can be used to assign a descriptive sentence to a stated image, or to gain pictures that prove a given sentence. The score is attained with the aid of evaluating an assist of which means acquired from the picture to one acquired from the sentence. Each approximation of meaning comes from a discriminative technique this is learned using information.

P. Felzenszwalb et al. [10] depicts a discriminatively organized, deformable element display for item detection this is multi-scale. This framework accomplishes a two-fold development in ordinary precision over the pleasant show in the 2006 PASCAL character acknowledgment venture. The framework depends vigorously on deformable parts. While deformable element models have grown to become out to be modestly well-known, their nice had no longer been installation on troublesome benchmarks, as an example, the PASCAL project. This machine also is predicated closely on new methods for discriminative schooling. They integrate a margin-sensitive approach for information mining difficult terrible samples with a formalism known as latent SVM. A latent SVM, like a shrouded CRF, activates a non-curved preparing difficulty. But, a latent SVM is semi-convex and the training difficulty converts curved as soon as latent information is exact for the fine examples. Authors agree with that their schooling methods will eventually make viable the effective use of greater latent data consisting of hierarchical (grammar) models and fashions regarding latent three-dimensional pose.

Y. Gotoh et al. [11] addressed generation of herbal language descriptions for human actions, conduct and their institutions with different things located in video streams. In this, they projected conventional picture processing techniques to extract excessive-degree a functions from a video. Those functions are altered into herbal language descriptions by using context-unfastened grammar.

I. Laptev et al. [13] tended to acknowledgment of natural human sports in differing and realistic video settings. This animating however vital problem has for the maximum component been neglected inside the past due to diverse issues considered one of that is the absence of affordable and commented on video datasets. Their first contribution is to cope with this restrict and to research the use of movie scripts for automated human moves annotation in videos. They assess elective techniques for activity recuperation from scripts and display focal factors of a content based classifier. The use of the retrieved movement examples for visible gaining knowledge of, they flip to the following problem of movement classification in a video. They introduce a singular method for video arrangement that expands upon and broadens some past due thoughts with space-time pyramids, community space-time highlights, and multichannel non-directly SVMs. They eventually follow the approach to gaining knowledge of and classifying idea and hectic movement classes in movies and display promising consequences.

Laptev et al. [14] proposed a model for semantic rationalization of events, just like weddings or b-ball games. The framework carries event taxonomy, applied as a faceted category, and an event paratomy, realistic the usage of the ABC ontology.

Lee et al. [15] recommend a high-stage picture instance, referred to as the item financial institution in which a photograph is indicated as a scale-invariant response map of great pre-skilled fashionable object locators, oblivious in regards to the trying out dataset or visible project. Siming Li et al. [16] gift a modest yet powerful approach to routinely compose photograph descriptions assumed pc vision primarily based inputs and the use of web-scale n-grams. A unique maximum preceding study that summarizes or recovers pre-existing text well sized to an image, their projected technique accommodates sentences totally from scratch.

Yuri Lin et al. [17] present a novel release of the Google Books Ngram Corpus that depicts how automatically phrases and expressions have been use over a time of five centuries, in eight dialects. This novel model provides syntactic comments, as an example, words are labeled with their grammatical form, and head-modifier affiliations are recorded. The annotations are made consequently thru real exhibitions that are precisely adapted to historic content material.

Tanvi and Mooney [18] gift a brand new mixture of trendy item recognition, hobby type, and textual content mining to study effective pastime recognizers deprived of ever definitely labelling schooling movies. They devise cluster verbs used to define films to robotically adjust lessons of activities and yield a labeled education set. This labeled data is then used to prepare an action classifier taking into consideration spatiotemporal elements. Second, text mining is introduced to examine the associations among those verbs as well as related items. This fact is then used with the outputs of an off-the-shelf object recognizer as well as the skilled hobby classifier to create a better activity recognizer.

Ben Packer et al. [19] presented a system that is able to recognize difficult, best-grained human moves with the control of objects in sincere motion sequences.

Kishore k. Reddy et al. [20] advice the scene context information obtained from transferring and motionless pixels within the key frames, in aggregate with motion features, to solve the motion popularity difficulty on a large dataset with movies from the internet. Heng Wang et al. [21] proposed a technique to define movies by manner of dense trajectories. Dense factors from every frame or picture inspected and tune them taking into account development data from a dense optical float subject. Trajectories are strong to short unpredictable moves and in addition shot impediments via giving a present day optical drift algorithm. Furthermore, dense routes defend the motion records in films well.

Yezhou Yang et al. [22] deliberate a sentence technology method that designates photos by way of forecasting a likely nouns, verbs, scenes and prepositions that form the core sentence shape. The enter is a noisy estimation of the objects and scenes detected in the body/photograph with a nation of the art trained detectors. They make use of those appraisals as parameters on a Hidden Markov model (HMM) that fashions the sentence era manner, with hidden nodes as selection elements and picture recognitions because the emanations.

B. Yao et al. [23] give object and human position because the context of each other in different Human item conversation (HOI) activity classes. They increase a random field version that makes use of a construction getting to know technique to analyze large connectivity patterns among gadgets and human body parts.

R. M. Kagalkar et al. [24] offered a method for detecting tumours in breast using template-matching. In [25] they used the implicit contour method for sufferers' X-ray image segmentation. R. Kagalkar and M. Patile [25] [26] introduced picture to textual content conversion (ITT) and speech to text (STT) the use of object reputation method.

R.Kagalkar and V.Edke [28] introduced video object description of short videos in Hindi text version and use of preprocessing, segmentation feature extraction techniques.

### **3. PROPOSED SYSTEM OVERVIEW**

The proposed technique can be considered as a holistic records-pushed three-step manner in which items are firstly detected and activities the use of latest visual reputation strategies. After that, this frequently noisy detection is part of with an assessment of actual-global opportunity, which accomplishes by mining SVO triplets from full-size scale net corpora. To the cease, these triplets are used to make candidate sentences which might be then ranked for reliability and grammaticality. Item detectors had been considered for static set of videos, each video became break up into frames at one-second intermissions. For each frame, the item detectors are run and selected the most rating assigned to respective item in any of the frames. In order to get an underlying prospect conveyance for activities recognized within the films, the motion descriptors are utilized. These descriptors are then randomly examined and clustered to benefit a “bag of seen phrases,” and every video is then denoted as a histogram over these clusters. The subject, verb and item from the pinnacle-scoring SVO are used to supply a hard and fast of candidate sentences, which can be then ranked the use of a language model. The advanced scheme changes phrases and sentences of Indian natural language into text in Hindi. The authority of video processing strategies and synthetic intelligence techniques has used to obtain the intention. To perform the errand effective photograph prepared systems are utilized, as an example, define differencing based totally monitoring, edge reputation, image fusion, to place shapes in movies. Facet popularity, picture fusion, to area shapes in movies.

#### **3.1. Product Overview**

The proposed system consists of two major module training and testing is shown in figure 1.

##### **3.1.1. Training Video:**

The training section is used to train videos and stored on the database with its features and SVO description which need for video testing.

- Firstly, the video is split into images or frames since a video is nothing but a set of images. Training is performed on short videos because frames of long video are more. If there are a number of images is additional than time require to process one video will more.
- After that, every Image is processed by purifying (noise removal, edge detection) and applying Scale-invariant feature transform (or SIFT) feature extraction algorithm.
- Triplets are used to produce candidate sentences which are then ranked for likelihood as well as grammaticality. This section is handled by the admin who is liable for data training.

##### **3.1.2. Testing Video:**

This module test video learner and gets the result if at slightest one video is trained.

- In this phase, a video is processed and divided into frames and these frames are further processed by applying the purifying algorithm to remove noise from images. Gaussian filtering technique is used to filter image.
- After elimination of noise, the features of images are extracted to detect objects.

- These features are linking with training videos to recognize Hindi text.

## 4. METHODOLOGY

### 4.1. Proposed System Architecture

The system architecture for proposed technique is elaborated in this section.

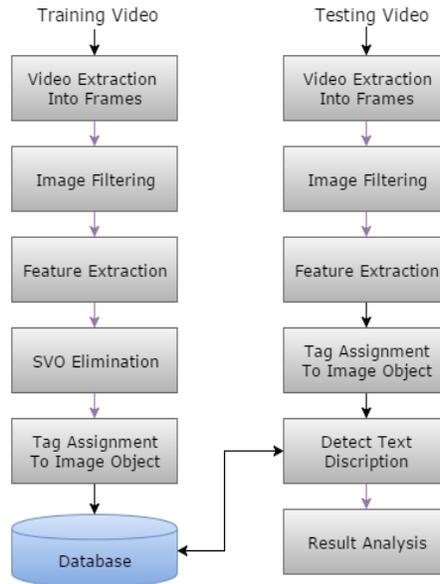


Figure 1. System Architecture

### 4.2. Overview

#### 4.2.1. Pre-process

This section consist pre-processing on video like noise and blur elimination and edge detection. Video holds huge amount of data at dissimilar levels in terms of sights, shots and surrounds. Gaussian filtering technique is used to eliminate blur from images and remove noise and detail. Graphically Gaussian distribution can see as bell shape if mean is 0 and standard deviation of the distribution  $\sigma=1$ . For working with images need to use the two dimensional Gaussians function. Gaussian filtering is more effective at smoothing images. It has its premise in the human visual recognition framework. It has been found that in the human visual discernment structure. It has been found that neurons make a comparative filter when handling visual images. Canny edge discovery procedure is utilized to recognize edges of items present in pictures or edges.

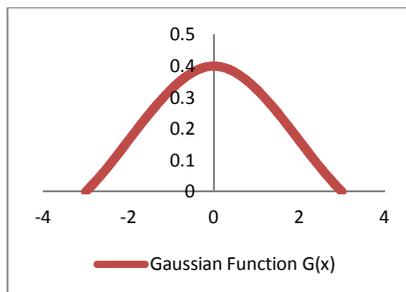


Figure 2. Gaussian distribution graph

#### 4.1.2. Segmentation

Segmentation partitions an image into particular areas containing every pixel with comparable attributes. To be significant and valuable for image examination and elucidation, the areas should unequivocally identify with depicted objects or components of interest. Meaningful segmentation is the initial step from low-level image handling changing a greyscale or shading image into one or more different images to abnormal state image depiction as far as elements, articles, and scenes.

#### 4.2.3. Feature Extraction

The SIFT approach, for image highlight era, takes an image and changes it into an "expansive collection of local feature vectors". Each of these feature vectors is invariant to any scaling, resolution or interpretation of the image. This methodology offers numerous components with neuron reactions in primate vision. To help the extraction of these elements the SIFT algorithm applies a 4 stage separating approach:

- 1) Scale-Space Extreme Detection
- 2) Key point Localization
- 3) Orientation Assignment
- 4) Key point Descriptor

#### 4.2.4 SVM Classification

SVM classification is essentially a binary (two-class) classification technique, which has to be modified to handle the multiclass tasks in real world situations. SVM classification uses features of image to classify.

#### 4.2.5 Neural Network Classification

The neural network classifier system is built for the features utilizing the back-proliferation learning classifier.

### 5. VIDEO PROCESSING

Input: V – Video (containing objects as well as events)

Process:

1. Convert Video into image frames.
2. Convert RGB (Red, Green, and Blue) colour video into Grayscale video by eliminating the hue and saturation information and retaining the luminance.
3. Apply Gaussian Filter for noise and blur elimination.

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

4. Apply Image segmentation by performing edge detection algorithm is proposed based on morphology, canny edge detector.
5. Apply SIFT Descriptor and extract image features.
6. For each frame, object and activity detectors are used to detect objects in a frame for selecting the most probable subject-verb-object triplet for describing a video.

7. Compute detection scores for each frame and converted the detection scores into the function to estimate probability using sigmoid Perform text mining using Stanford dependency parser.
8. Estimating SVO probabilities.
9. Compute similarity between Verbs using WUP similarity. WUP similarity between the original ( $V_{orig}$ ) and expanded ( $V_{sim}$ ) verbs can compute as:

$$score = w_1 \times vis_{score} + w_2 \times nlp_{score},$$

Where,  $vis_{score} = P(S|vid) \times P(V_{orig}|vid) \times Sim(V_{sim}, V_{orig}) \times p(O|vid)$

Where, P denoted probability, S denotes subject, V denotes verb and O denotes object.

10. When computing the overall vision score, make a conditional independence guess and multiply the likelihoods of the subject, activity, and object.
11. Lastly, the subject, verb and object from the top-scoring SVO are used to create a set of contender sentences, which are then ranked using a language model
12. Compare features and analyse result.

## 6. EXPERIMENT AND ANALYSIS

### 6.1. Data Description

The dataset is made of Hindi word description of videos that are used as training part. These videos are divided into various categories like sports, animals, rural area, urban area, hospitality, Martian area, natural scene, collage area, airplane, etc.

Table 1. Database Description.

Video	Category	Description in English	Description in Hindi
	पुष्ट वीडियो (Athletic video)	In this video a man is riding on a bicycle and a man running behind the bicycle.	इस वीडियो में एक आदमी को एक साइकिल की सवारी कर रहा है और एक आदमी है कि साइकिल आदमी के साथ भाग रहा है
	खेल (Sport) लम्बी कूद (Long Jump)	It is a long jump video. In this video one woman is playing a long jump.	यह लंबी कूद का वीडियो है इस वीडियो में एक महिला लंबी कूद खेल रही है वह मंच पर चल रही है और रेत पर कूद रही है
	खेल (Sport) फुटबॉल (Football)	It is the video of football. In this video two players are playing the football.	इस फुटबॉल का खेल है। इस दो खिलाड़ियों में खेल रहे हैं एक खिलाड़ी फुटबॉल के साथ खेल रहा है और किक करने के लिए जा रहा

	<p>जानवर (Animal)</p>	<p>In this video one small child is playing with the dog and the small child is smiling.</p>	<p>इस वीडियो में एक बच्चे को एक कुत्ते के साथ खेल रहा है। कुत्ते खुशी से बुलबुला खेल रहा है और बच्चे मुस्कुरा रही है</p>
	<p>अस्पताल (Hospital)</p>	<p>It is the video of hospital. Patient is sleeping on a bed. One woman is sitting near the patient. Doctor is standing near the patient.</p>	<p>यह वीडियो अस्पताल की है। इस में एक रोगी के बिस्तर पर है। और एक महिला रोगी के पास बैठा है। लेडी डॉक्टर मरीज को सलाह दे रहा है।</p>
	<p>महाविद्यालय के पुस्तकालय (Collage Library)</p>	<p>It is the video of college library. In this video there is one girl and one boy. They are reading the books.</p>	<p>इस वीडियो में एक लड़की और एक लड़का महाविद्यालय के पुस्तकालय में किताबें पढ़ रहे हैं।</p>
	<p>महाविद्यालय के पुस्तकालय (Collage Library)</p>	<p>It is the video of college library. In this video one girl is writing on a notebook.</p>	<p>यह वीडियो महाविद्यालय के पुस्तकालय की है। इस वीडियो में एक लड़की एक कलम के साथ नोटबुक पर लिख रही है। उसकी बेंच पर कई किताबें हैं।</p>
	<p>हवाई जहाज़ (Airplane)</p>	<p>It is video of aeroplane. Aeroplane is just going to take off.</p>	<p>इस वीडियो में हवाई जहाज रनवे पर चलाने के लिए शुरू है। कुछ समय के बाद हवाई जहाज उड़ान भरने के लिए जा रहा है।</p>

## 6.2. Results and Description

Some likely results are predicted using some minor length of videos as shown in the table given below.

Table 2. Prediction Output.

Video	Dependent Output in English	Dependent Output in Hindi	Description
	Girl riding on a horse.	बाल घोड़े पर सवारी कर रहा	बाल घोड़े पर सवारी कर रहा है
	Boy is riding on a bike.	आदमी सड़क पर एक साइकिल सवारी कर रहा है। उनकी गति बहुत तेज है	आदमी सड़क पर एक साइकिल की सवारी कर रहा है। उनकी गति बहुत तेज है
	Dog is running.	कुत्ते बहुत तेजी से भाग रहा है	कुत्ता बहुत तेजी से भाग रहा है

### 6.3. Table Values

Table 3 shows the object predicted from some video of different category. This may be give less result to find objects on testing video. The video is tested according to trained video but some part of video is removed before test to evaluate result.

Table 3. Predicted Objects.

Category	Actual Objects	Predicted Objects	Percentage (%)
खेल (Sport)	2	2	100
महाविद्यालय (collage)	6	5	83
अस्पताल (Hospital)	8	6	75
हवाई जहाज़ (Airplane)	3	3	100
जानवर (Animal)	3	2	67

## 7. CONCLUSIONS

This paper has introduced a brand new method for producing Hindi language descriptions of videos via classifying the great subject-verb-object triplet for describing sensible motion pictures. The object detectors had been deliberate for static images therefore, every video splits into frames at one-second duration and the object detectors are implemented on each frame. Features are mined using SIFT set of rules and those features are used for contrast of trying out with the training video. In future, system will construct the gadget that generates text description for greater complicated sentences with adjectives, adverbs, and more than one gadgets and multi-sentential descriptions of longer motion videos with a couple of activities.

## REFERENCES

- [1] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, Siming Li, Y. Choi, A. C. Berg, and Tamara L. Berg, "BabyTalk: Understanding and Generating Simple Image Descriptions", *IEEE Trans on pattern analysis and machine intelligence*, vol. 35, no. 12, Dec 2013.
- [2] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, and S. Guadarrama, "Generating Natural-Language Video Descriptions Using Text-Mined Knowledge", 2013
- [3] Vandana D. Edke and Ramesh M. Kagalkar, "Review Paper on Video Content Analysis into Text Description", *International Journal of Computer Applications National Conference on Advances in Computing (NCAC-2015)*, 2015.
- [4] Rashmi B. Hiremath and Ramesh M. Kagalkar, "Review Paper on Sign Language Recognition Techniques", *International Journal of Computer Applications National Conference on Advances in Computing (NCAC 2015)*, 2015.
- [5] Barbu, A. Bridge, A. Burchill, Z. Coroian, D. Dickinson, S. Fidler, S. Michaux, A. Mussman, S. Narayanaswamy, S. Salvi, et al., "Video in sentences out", In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 102–12, 2012.
- [6] Chang, C., and Lin, "LIBSVM: a library for support vector machines", *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27, 2011.
- [7] De Marneffe, M. MacCartney, B. and Manning, "Generating typed dependency parses from phrase structure parses", In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 6, 449–454, 2006.
- [8] Ding, D. Metze, F. Rawat, S. Schulam, P. Burger, S. Younessian, E. Bao, L. Christel, M. and Hauptmann, "Beyond audio and video retrieval: towards multimedia summarization", In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 2012.
- [9] Farhadi, A. Hejrati, M. Sadeghi, M. Young, P. Rashtchian, C. Hockenmaier, J. and Forsyth, D., "Every picture tells a story: Generating sentences from images," *Computer Vision–European Conference on Computer Vision (ECCV)* 15–29, 2010.
- [10] Felzenszwalb, P. McAllester, D. and Ramanan, D., "A discriminatively trained, multiscale, deformable part model", In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8, 2008.
- [11] Khan, M. U. G., and Gotoh, Y., "Describing video contents in natural language", In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, 27–35. Association for Computational Linguistics, 2012.
- [12] Kulkarni, G. Premraj, V. Dhar, S. Li, S. Choi, Y. Berg, A. and Berg, T., "Baby talk: Understanding and generating simple image descriptions", In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1601–1608, 2011.
- [13] Laptev, I., and Perez, P., "Retrieving actions in movies", In *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV)*, 1–8, 2007.

- [14] 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), 1–8, 2008.
- [15] Lee, M. Hakeem, A.; Haering, N. and Zhu, S., “Save: A framework for semantic annotation of visual events”, In IEEE Computer Vision and Pattern Recognition Workshops (CVPR-W), 1–8, 2008.
- [16] Li, S. Kulkarni, G. Berg, T. Berg, A. and Choi, Y., “Composing simple image descriptions using web-scale n-grams”, In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL), 220–228, Association for Computational Linguistics (ACL), 2011.
- [17] Lin, Y. Michel, J. Aiden, E. Orwant, J. Brockman, W. and Petrov, S., “Syntactic annotations for the google books ngram corpus”, In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), 2012.
- [18] Motwani, T., and Mooney, R., “Improving video activity recognition using object recognition and text mining, European Conference on Artificial Intelligence (ECAI), 2012.
- [19] Packer, B.; Saenko, K.; and Koller, D., “A combined pose, object, and feature model for action understanding”, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1378–1385, 2012.
- [20] Reddy, K., and Shah, M., “Recognizing 50 human action categories of web video”, Machine Vision and Applications 1–11, 2012.
- [21] Wang, H.; Klaser, A.; Schmid, C.; and Liu, C.-L., “Action recognition by dense trajectories”, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3169–3176, 2011.
- [22] Yang, Y. Teo, C. L. Daume, III, H. and Aloimonos, Y., “Corpus-guided sentence generation of natural images”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 444–454, Association for Computational Linguistics, 2011.
- [23] Yao, B., and Fei-Fei, L., “Modeling mutual context of object and human pose in human-object interaction activities”, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.
- [24] Ramesh. M. Kagalkar, Mrityunjaya V. Latte and Basavaraj M. Kagalkar “Template Matching Method For Localization Of Suspicious Area And Classification Of Benign Or Malignant Tumors Area In Mammograms”, International Journal on Computer Science and Information Technology (IJCECA), ISSN 0974-2034, Vol.25, Issue1, 2011.
- [25] Ramesh M. Kagalkar Mrityunjaya .V. Latte and Basavaraj. M. Kagalkar ““An Improvement In Stopping Force Level Set Based Image Segmentation”, International Journal on Computer Science and Information Technology(IJCEIT), ISSN 0974-2034,Vol 25,Issue1,Page 11-18,2010.
- [26] Mrunmayee Patil and Ramesh Kagalkar “An Automatic Approach for Translating Simple Images into Text Descriptions and Speech for Visually Impaired People”, International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 3, May 2015.
- [27] M. Patil and Ramesh Kagalkar “A Review on Conversion of Image to Text As well As Speech Using Edge Detection and Image Segmentation” International Journal of Advance Research in Computer Science Management Studies, Volume 2, and Issue 11 (November-2014) publish on 29th November to 30th November 2014.
- [28] Vandana Edke and Ramesh M. Kagalkar“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

## Authors

**Vandana D. Edke:** is M.E 2nd year student of Computer Engineering Department, Dr. DY Patil School of Engineering and Technology, Lohegaon, Pune .



**Ramesh. M. Kagalkar** was born on Jun 1st, 1979 in Karnataka, India and presently working as an Assistant. Professor, Department of Computer Engineering, Dr. D. Y. Patil School Of Engineering and Technology, Charoli, B.K.Via –Lohegaon, Pune, Maharashtra, India. He has 13.5 years of teaching experience at various institutions. He is a Research Scholar in Visveswaraiah Technological University, Belgaum, He had obtained M.Tech (CSE) Degree in 2006 from VTU Belgaum and He received BE (CSE) Degree in 2001 from Gulbarga University, Gulbarga. He is the author of text book Advance Computer Architecture which covers the syllabus of final year computer science and engineering, Visveswaraiah Technological University, Belgaum. One of his research article “A Novel Approach for Privacy Preserving” has been consider as text in LAP LAMBERT Academic Publishing, Germany (Available in online). He is waiting for submission of two research articles for patent right. He has published more than 36 research papers in International Journals and presented few of there in international conferences. His main research interest includes Image processing, Gesture recognition, speech processing, voice to sign language and CBIR. Under his guidance four ME students awarded degree in SPPU, Pune, five students at the edge of completion their ME final dissertation reports and two students started are started new research work and they have publish their research papers on International Journals and International conference.

