International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 12, Number 2 (2016), pp. 103-116

http://www.ripublication.com

# Video Object Description of Short Videos in Hindi Text Language

**Vandana D. Edke and Ramesh M. Kagalkar**

*2nd Year M.E Student of Dept. of Computer Engineering.*
*Dr. DY Patil School of Engineering. and Technology, Pune*
*Email: vandana.edke@gmail.com*
*Research Scholar and Asst. Professor, Dept. of Computer Engineering.*
*Dr. D Y Patil School of Engineering and Technology, Pune*
*Email: ramesh.kagalkar@dypic.in*

**Abstract**

Video object recognition has become a creative area of analysis in recent years. This strategy combines the outcome of state-of-the-art object and activity indicators with "real-world" information to pick the in all subject-verb-object triplet for depicting a video. A general data-driven approach is denoted in this paper that produces descriptions of video content into text description in the Hindi language. The proposed work provides preliminary and basic text description in the Hindi language that is producing simple words and sentence formation. But the main challenging effort in this work is to extract grammatically correct and expressive text information in Hindi text regarding video content. Using this triplet choice technique, a video is tagged by the trainer, in particular, Subject, Verb, and Object (SVO) and then this data is mined to improve the result of testing video explanation by means of activity as well as object identification.

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

## 1. Introduction

Video processing is one of the majority increasing field in research and technology in today's world. Increasing sharing of public photo and video on websites, such as "Flickr" and "YouTube" delivers a massive corpus of unstructured video and image data over the Internet. Retrieving visual evidence from the Web, however, has been regularly limited to the use of meta-data, user-annotated tags, subtitles and

surrounding text (E.g. the image search engine used by Google [1]). Combining natural-language processing (NLP) with computer vision to generate Hindi descriptions of visual data is a significant area of active research. The paper presents a novel method for describing short videos. First, recognize the most probable subject, verb and object triplet for identifying visual item and activity detectors and text mined information to access the probability of SVO triplets Secondly, assumed the selected SVO triplet; it uses a simple template based method to generate candidate sentences which are then ranked by means of a statistical language model prepared on web-scale data to get the best worldwide portrayal. This is surface realization stage.

The proposed method can be viewed as a three-step process where objects are firstly detected and activities using state-of-the-art visual recognition techniques. Besides, consolidate these frequently noisy detections with an evaluation of true likelihood, which is obtain by mining SVO triplets from extensive scale web corpora. The subsequent natural-language descriptions can be usefully employed in claims such as semantic video search and summarization and providing video interpretations for the outwardly impeded. Finally, these triplets are used to create candidate sentences which are then ranked for plausibility as well as grammaticality [2].

The presented method is complex in annotating arbitrary short videos using off-the-shelf visual detectors, deprived of the engineering effort essential to building domain-specific activity models. The main contribution is including the pragmatics of various entities' probability of being the subject/object of a given activity, learned from web-scale text corpora. For instance, animate objects like people and animals are more likely to be subjects likened to inorganic objects like balls or TV monitors. Equally, certain objects are more likely to function as subjects or objects of definite activities, e.g., "riding a horse" verses "riding a house.

Picking the best verb may likewise key perceiving activities for which no unequivocal training information has been conveyed. For instance, think about a video with a woman walking her pet cat. The object locators may recognize the woman and the cat; yet the activity indicators may basically have the more general movement, "move" in their training data. In such circumstances, real world pragmatics is very useful in signifying that "walk" is best used to designate a woman "moving" with her cat. This process is referring as verb expansion. Finally, results are computed by using real world short length videos.

In this paper, a description of video content into text using Hindi language is proposed. In section 1, related work on video text description and objects detection from video is mentioned. In section 2 Literature survey is carried out and in section 3 depicts the proposed framework review and In section 4 demonstrates the Methodology for proposed system. In section 5 describes the dataset and predicted results. And finally, concludes the paper.

## 2. Literature Survey

In the literature review, we are going to debate topical methods over the video text recognition: Below in literature we are debating some of them.

V. Edke and R. Kagalkar [3] described review on video content study into text description. Thus this paper presented three necessary contributions to activity recognition from video. Firstly, they introduced a single mechanism for automatically discovering videos activity categories from natural-language descriptions. Secondly, an existing activity recognition scheme is improved abuse object context along with relationships between objects and activities. Finally, shows language process is familiar automatically extracting the requisite data about the relationship between objects and activities from a corpus of general text.

R. Hiremath and R. Kagalkar [4] presented review on sign language recognition for the key finding of the comparative analysis of similar techniques and also for technology used in vision based hand gesture recognition. Ding, D. et al. [5] review preceding study on audio as well as video processing, and describe the Topic-Oriented Multimedia Summarization (TOMS) task using Natural Language Generation: given a set of automatically mined features from a video. A Topic-Oriented Multimedia Summarization (TOMS) framework will consequently create a passage of common dialect, which outlines the critical information in a video having a place with a specific point range, and gives elucidations to why a video was coordinated, recovered, and so forth. They show this as a first stage towards schemes that will be able to discriminate visually similar, but semantically different videos, associate two videos and give composed yield or summarize a substantial number of videos without a moment's delay. Authors present methodology of determining the TOMS issue. They remove Visual Concept components and ASR interpretation and enhance a Template-Based Natural Language Generation (NLG) Scheme to create a composed relating in view of the mined elements. Authors additionally propose conceivable designs plans for continuously assessing and refining TOMS frameworks, and present consequences of a pilot designs of initial framework [5].

De Marneffe [6] depicts a framework for removing typed reliance parses of English sentences since expression structure parses. So as to capture basic relations going on in corpus texts that can be unsafe in real-world applications, numerous Noun Phrase (NP) relations are included in the set of grammatical relations used.

Chang, C. et al. [7] presented complete implementation details of Support Vector Machines called LIBSVM. Though, this article does not a goal to explain the practical use of LIBSVM for guidelines of using LIBSVM.

P. Felzenszwalb et al. [8] depicts a discriminatively prepared, deformable part display for item detection that is multi-scale. This framework accomplishes a two-fold development in normal precision over the best show in the 2006 PASCAL individual acknowledgment challenge. The framework depends vigorously on deformable parts. While deformable part models have turned out to be modestly famous, their quality had not been set up on troublesome benchmarks, for example, the PASCAL challenge. This system also relies heavily on new approaches for discriminative training. They combine a margin-sensitive method for data mining tough negative samples with a formalism called as latent SVM. A latent SVM, like a shrouded CRF, prompts a non-curved preparing issue. However, a latent SVM is semi-convex and the training difficulty converts curved once latent information is specified for the positive examples. Authors believe that their training methods will finally make possible the

effective use of more latent info such as hierarchical (grammar) models and models involving latent three-dimensional pose.

Ali Farhadi [9] et al. defined a system that can compute a score involving an image to a sentence. This score can be used to assign a descriptive sentence to a stated image, or to gain images that prove a given sentence. The score is attained by comparing an assist of meaning obtained from the image to one obtained from the sentence. Each approximation of meaning comes from a discriminative process that is learned using data.

Y. Gotoh et al. [10] addressed generation of natural language descriptions for human actions, behavior and their associations with other things observed in video streams. In this, they projected conventional image processing methods to extract high-level a features from a video. These features are altered into natural language descriptions by means of context-free grammar.

Laptev et al. [11] proposed a model for semantic explanation of occasions, similar to weddings or b-ball games. The framework contains event taxonomy, applied as a faceted classification, and an event partonomy, practical using the ABC ontology.

Laptev et al. [12] tended to acknowledgment of natural human exercises in differing and practical video settings. This animating however vital subject has for the most part been disregarded in the past because of various issues one of which is the absence of reasonable and commented on video datasets. Their first contribution is to address this restriction and to investigate the use of movie scripts for automatic human actions annotation in videos. They assess elective strategies for activity recovery from scripts and show focal points of a content based classifier. Using the retrieved action examples for visual learning, they turn to the following problem of action classification in a video. They introduce a novel strategy for video arrangement that expands upon and broadens a few late thoughts with space-time pyramids, neighborhood space-time highlights, and multichannel non-straight SVMs. They finally apply the method to learning and classifying thought and irritating action classes in movies and show promising results.

Lee et al. [13] propose a high-level image illustration, known as the Object Bank in which an image is indicated as a scale-invariant response map of enormous pre-trained general object locators, oblivious in regards to the testing dataset or visual task.

Yuri Lin et al. [14] present a novel release of the Google Books Ngram Corpus that depicts how routinely words and expressions were use over a time of five centuries, in eight dialects. This novel version presents syntactic remarks, for example, words are labeled with their grammatical form, and head-modifier affiliations are recorded. The annotations are made consequently through factual exhibitions that are precisely adapted to historical content.

Siming Li et al. [15] present a modest yet effective method to automatically compose image descriptions assumed computer vision based inputs and using web-scale n-grams. A different most previous study that summarizes or recovers pre-existing text significant to an image, their projected method comprises sentences entirely from scratch.

Tanvi and Mooney [16] present a new mixture of standard object recognition, activity classification, and text mining to study effective activity recognizers deprived of ever clearly labeling training videos. They create cluster verbs used to define videos to automatically regulate classes of activities and yield a labeled training set. This labeled information is then used to prepare an action classifier taking into account spatiotemporal elements. Second, text mining is added to learn the associations among these verbs as well as related objects. This information is then used with the outputs of an off-the-shelf object recognizer as well as the trained activity classifier to create a better activity recognizer.

Ben Packer et al. [17] presented a system that is able to recognize difficult, fine-grained human actions with the management of objects in truthful action sequences.

Heng Wang et al. [18] proposed a method to define videos by dense trajectories. Dense points from every frame or image inspected and track them taking into account development information from a dense optical flow field. Trajectories are robust to quick unpredictable movements and in addition shot impediments by giving a state-of-the-art optical flow algorithm. Moreover, dense routes shield the motion information in videos well.

Kishore K. Reddy et al. [19] propose the scene context info obtained from moving and immobile pixels in the key frames, in combination with motion features, to resolve the action recognition difficulty on a big dataset with videos from the web.

Yezhou Yang et al. [20] planned a sentence generation approach that designates images by forecasting a possible nouns, verbs, scenes and prepositions that form the core sentence structure. The input is a noisy estimation of the items and scenes detected in the frame/image with a state of the art trained detectors. They utilize these appraisals as parameters on a Hidden Markov Model (HMM) that models the sentence generation process, with hidden nodes as decision parts and picture recognitions as the emanations.

R. M. Kagalkar et al. [21] presented an approach for detecting tumors in breast using template-matching. In [25] they used the implicit contour method for patients' X-ray images segmentation. R. Kagalkar and M. Patil [22] introduced image to text conversion (ITT) and speech to text (STT) using object recognition technique.

## 3. Proposed System

The proposed method can be viewed as a holistic data-driven three-step process where objects are firstly detected and activities using state-of-the-art visual recognition techniques. After that, these frequently noisy detections join with an assessment of real-world probability, which accomplish by mining SVO triplets from substantial scale web corpora. To the end, these triplets are used to make candidate sentences that are then ranked for reliability and grammaticality. Discriminatively-trained [23] deformable portions representations used to detect the maximum probable things in each video. Since these object detectors were considered for static set of images, every video was split into frames at one-second intermissions. For every frame, the object detectors are run and selected the maximum score assigned to respective object in any of the frames. So as to get an underlying prospect conveyance for activities

recognized in the videos, the movement descriptors are utilized. These descriptors are then randomly tested and clustered to achieve a "bag of visual words," and each video is then denoted as a histogram over these clusters. The subject, verb and object from the top-scoring SVO are used to produce a set of candidate sentences, which are then ranked using a language model. The developed scheme changes words and sentences of Indian Natural language into text in Hindi. The authority of image processing techniques and artificial intelligence techniques has used to attain the objective. To accomplish the errand effective picture preparing systems are utilized, for example, outline differencing based tracking, edge recognition, image fusion, to area shapes in videos[24].

The proposed system presents a holistic data-driven methodology for generating Hindi language descriptions of short videos by identifying the best subject-verb-object triplet for describing videos. The proposed system consists of two major module training and testing is shown in figure 1.

**1) Training Module:**

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 [25].

**2) Testing Module:**

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 Preprocess

This section consist preprocessing on video like frame extraction, noise and blur elimination and edge detection. Video holds huge amount of data at dissimilar levels in terms of sights, shots and surrounds. Thus to process on video, first extract frames from video. These frames are nothing but images that are used for further processing.

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. This is simply the product of two 1D and 2D Gaussian functions. 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 [26]. The Canny calculation essentially discovers edges where the grayscale e-intensity of the picture changes the most. These areas are found by deciding angles of the picture. Gradients at every pixel in the smoothed picture are controlled by applying what is known as the Sobel-operator. The gradient magnitudes (otherwise called the edge strengths) can then be resolved as a Euclidean distance measure by applying the law of Pythagoras.

**4.2 Segmentation**

Segmentation partitions an image into particular areas containing every pixel with comparable attributes. Image segmentation is employed for understanding image contents. It is terribly tough to know the advanced images and videos hence image segmentation is employed for analyzing the content or smaller components of image. Using image segmentation object detection ought to be takes place. Object detection is a crucial half for recognizing image content.

**4.3 Feature Extraction**

For any object there are numerous elements, interesting points on the object that can be extricated to give a "feature" description of the object. This description used when attempting to locate the object in an image containing many other objects. The SIFT approach, for image highlight era, takes an image and changes it into an "expansive collection of local feature vectors". This methodology offers numerous components with neuron reactions in primate vision [27]. To help the extraction of these elements the SIFT algorithm applies a 4 stage separating approach:

1) Scale-Space Extrema Detection
2) Keypoint Localization
3) Orientation Assignment
4) Keypoint Descriptor

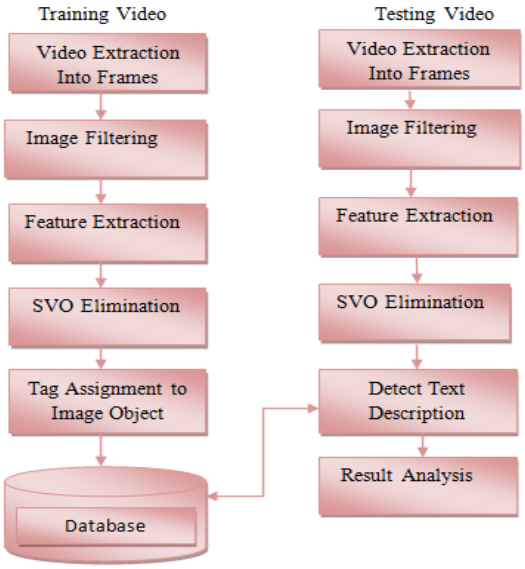


**Figure 1:** System Architecture

### 4.4 SVM Classification

SVM classification is essentially a binary (two-class) classification technique, it handles multiclass tasks in real world situations. SVM classification uses features of image to classify.

### Steps for algorithm Implementation

Input:

V – Video (containing objects as well as events)

Process:

1. Convert Video into image frames.
2. Convert RBG (Red, Green, and Blue) color 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 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:

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

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

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 analyze result.

## 5. Experimental Result Analysis

### 5.1 Database Information

In order to evaluate Hindi text extraction process from videos, at least, 100 videos are used to train and store into database. In the presented scenario, the testing video is before trained so that the presented tests previously include some video description to get the best result. The dataset is made of Hindi word description of images that are used as training [[28]. 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：** Dataset description.

| Video | Category | Description |
|---|---|---|
|  | खेल (Sport) फुटबॉल (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 |
|  | जानवर (Animal) | 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 |

| | | |
|---|---|---|
|  | अस्पताल (Hospital) | 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 । |
|  | महाविद्यालय के पुस्तकालय (Collage Library) | इस वीडियो में एक लड़की और एक लड़का महाविद्यालय के पुस्तकालय में किताबें पढ़ रहे हैं। |
|  | हवाई जहाज़ (Airplane) | 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 । |

**5.2 Results and Discussion**

Proposed system application is based on dependent approach and according to depending approach we get 100 percent result of testing videos.

**Table 1:** Prediction Output

| Video Samples | Dependent Output | Description |
|---|---|---|
|  | बाल घोड़े पर सवारी कर रहा है | बाल घोड़े पर सवारी कर रहा है |
|  | 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 |
|  | कुत्ता बहुत तेजी से भाग रहा है | कुत्ता बहुत तेजी से भाग रहा है |

### 5.3 Table Values

The proposed system tested for depending approach and according to depending approach and it gives 100 % result of testing videos. This result is computed by using number of objects present in video. System firstly train videos by inserting objects, activities and description into the database and tested some of them videos to evaluate result in that case we get 100% result for testing video as shown in table 3.

**Table 3:** Predicted objects.

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

### 5.4 Graph

The proposed system time require to process video depend on size of video is discussed and shown in figure 2. Require time for Gaussian filtering, Canny Edge detection and SIFT feature extraction.

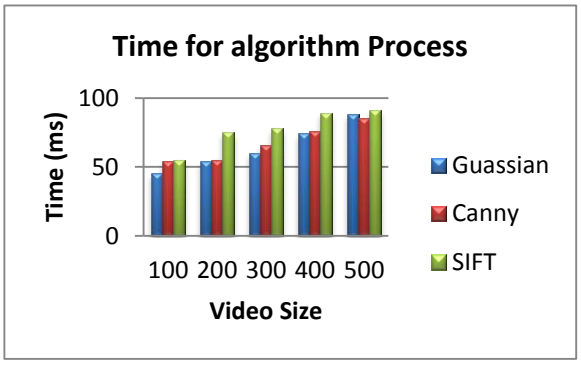


**Figure 2.** Shows time require to process video

### Conclusion

This paper has introduced a holistic data-driven method for generating Hindi language descriptions of short videos by classifying the best subject-verb-object triplet for describing realistic videos. This uses object detection, text mining, activity recognition and feature extraction. Each video splits into frames at one-second

intervals and the object detectors are applied on every frame. Features are mined using SIFT algorithm and these features ae used for comparison of testing with the training video. In future work try to build the system that generates text description for more complex sentences with adjectives, adverbs, and multiple objects and multi-sentential descriptions of longer videos with multiple activities.

**Author Biography**

**Vandana D. Edke**

She is M.E 2nd year student of Dept. of Computer Engineering, Dr. D Y Patil School of Engineering and Technology, Lohegaon, Pune. Her main research interest includes Image processing and Gesture recognition.

**Ramesh. M. Kagalkar**

He 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 14 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, One of his research article A Novel Approach for Privacy Pre-serving 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 35 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 Ten 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. He can be contacted by email rameshvtu10@gmail.com.

**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] 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.

[6] 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.

[7] Chang, C., and Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (TIST) 2(3):27, 2011.

[8] 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.

[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] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] Motwani, T., and Mooney, R., "Improving video activity recognition using object recognition and text mining, European Conference on Artificial Intelligence (ECAI), 2012.

[17] 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.

[18] 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.

[19] Reddy, K., and Shah, M., "Recognizing 50 human action categories of web video", Machine Vision and Applications 1–11, 2012.

[20] 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.

[21] 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.

[22] 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.

[23] 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 (IJCA), Volume 118, No. 3, May 2015.

[24] Mrunmayee and Ramesh Kagalkar, "A Review On Conversion of Image To Text as Well as Speech using Edge Detection and Image Segmentation", International Journal of Science and Research (IJSR), Volume 3, Issue 11, November 2014.

[25] Kaveri Kamble and Ramesh Kagalkar, "A Review: Translation of Text to Speech Conversion for Hindi Language", International Journal of Science and Research (IJSR), Volume 3, Issue 11, November 2014.

[26] Kaveri Kamble and Ramesh Kagalkar, "Audio Visual Speech Synthesis and Speech Recognition for Hindi Language",International Journal of Computer Science and Information Technologies (IJCSIT), CiiT International Journal of Data Mining Knowledge Engineering, Volume 6, Issue 2, April 2015.

[27] Kaveri Kamble and Ramesh Kagalkar " A Novel Approach for Hindi Text Description to Speech and Expressive Speech Synthesis" International Journal of Applied Information Systems (IJAIS), Volume 8, No.7, May 2015.

[28] Ramesh M. Kagalkar and Dr. S.V Gumaste, "Automatic Graph Based Clustering for Image Searching and Retrieval from Database", CiiT International Journal of Software Engineering and Technology, Volume 8, No 2, 2016.