

SUPPORT VECTOR MACHINE BASED APPROACH FOR TRANSLATING VIDEO SCENERIES TO NATURAL LANGUAGE DESCRIPTIONS

Vishakha Wankhede and Ramesh M. Kagalkar

Department of Computer Engineering, Dr. D Y Patil School of Engineering and Technology, India

Abstract

Human uses communication language either by written, spoken or typed to describe visual the world around them. So, the study of text description for any video goes increasing. This paper represents a framework that gives output as a description for any video having a maximum size of 50 seconds by using natural language processing. The framework is divided into two sections called training and testing. The training section is used to train the video with its description like activities of objects present in that video. The trained data is stored into the database with its features of scenario of video. Another section is testing section. The testing section is used to test the video and retrieve the output as description of video. By using Natural language processing sentences are generated from objects and their activities present in the video.

Keywords:

Natural language Processing, Video Processing, Video Recognition

1. INTRODUCTION

A combination of Natural-language processing (NLP) with computer vision is used to create text descriptions of visual information in a significant area. Generating natural language descriptions of visual content is an interesting task but involves combining the important research difficulties of visual recognition and natural language generation (NLG). While for explanations of images, current approaches have projected to statistically model the conversion from images to text. Visually descriptive language is a challenge for computer vision by joining imagery that is becoming additional related as recognition in addition to detection methods [1].

Computer vision has progressive to detect people, classify their activities, or to distinguish among many objects and specify their attributes. The output is a semantic representation encoding activities and objects categories. Whereas such representations of objects and activities can be well processed by automated systems, the common way to connect this information with individuals is natural language processing. Thus, this work addresses the problem of generating textual descriptions for videos. This work has a wide range of applications in the domain of human-computer/robot interaction, generating summary descriptions of videos, and automating movie descriptions for visually impaired person [2].

Human actions are (typically) defined by their appearances/motion characteristics and the complex and structured fundamental dependencies that relate them. These fundamental dependencies describe the goals and intentions of the agents. The storyline of a video includes the actions that occur in those video and causal relationships between them. A model that represents the set of storylines that can occur in a video corpus and the general causal relationships amongst actions in the video corpus is mentioned to as a "storyline model". The storyline model indicates the agents likely to achieve various actions and the visualization of

actions. A storyline model can be viewed as a (stochastic) grammar, whose language (individual storylines) signifies possible plausible "explanations" of novel videos in a domain. To reduce human labor, one can exploit the weak supervisory data in descriptions such as sportscaster interpretation. Many researchers have proposed using closed descriptions or other linguistic data to enhance video retrieval, video classification, or speech recognition.

In this paper, a method of generating textual description is proposed which explains human behavior appeared on real video by extracting semantic features of human motions and actions. In this paper, a description of video objects and activities into text is proposed. In section 2, related work on video text description and object detection from a video is mentioned. Section 3 depicts the proposed system overview. Section 4 demonstrates the methodology of the proposed framework. Section 5 gives a stepwise description of the video process. Section 6 describes the dataset and predicted results. And finally, section 7 concludes the paper.

2. RELATED WORK

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.

Bridge et al. [3] projected a scheme that yields sentential descriptions of video. The description of video contains who (object in video) did what to whom, and where and how they did it. Action of video class is extracted as a verb, member objects as noun phrases, properties of individual's objects as adjectival modifiers in those noun phrases, spatial associations among those contributors as prepositional phrases, and features of the occurrence as prepositional phrase adjuncts in addition to adverbial modifiers. Rohrbach et al. [4] intends to study the conversion from visual content to usual explanations from a parallel corpus of videos as well as textual descriptions instead of using rules in addition to templates to generate language adopting methods from statistical mechanism translation.

Gupta and Mooney [5] search how secure captions that certainly accompany numerous videos can act as weak supervision that permits automatically collecting 'labeled' information for activity recognition. In addition, authors propose caption classifier which uses extra linguistic data to decide whether a detailed comment denotes to an ongoing activity. Chang et al. [6] presented the overall implementation of Support Vector Machines known as LIBSVM. However, this complete article does not aim to describe the practical use of LIBSVM for strategies of using LIBSVM.

Marneffe et al. [7] depicted a structure for eliminating typed dependence parses of English language sentences since expression arrangement parses. So as to capture basic associations going on in corpus texts that can be unsafe in real-world applications, numerous Noun Phrase (NP) associations are included in the set of

grammatical relations used. Ding et al. [8] review study processing on audio as well as video, and describe the Topic-Oriented Multimedia Summarization (TOMS) task using Natural Language Generation (NLG). A Topic-Oriented Multimedia Summarization (TOMS) structure will, therefore, create a passage of common dialect, which plans the important information in a video having a place with a specific point range, and delivers elucidations for why a video was synchronized, recovered, and so forth. Authors moreover suggest conceivable strategy plans for continually assessing and refining TOMS frameworks and present consequences of a pilot designs of an initial framework. Farhadi [9] et al. defined a scheme that can calculate a score containing an image to a text sentence. This calculated score can be used to assign a descriptive sentence to a specified image or to obtain images that demonstrate a given sentence. The score is achieved by comparing an assistance of meaning achieved from the image to one gained from the sentence.

Felzenszwalb et al. [10] depicts a discriminatively prepared, deformable part display for item detection that is multi-scale. 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. They combine a margin-sensitive technique 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 curves once latent information is specified for the positive examples.

Gotoh et al. [11] addressed generation of natural language descriptions of human actions, behavior and their relations with other things detected in video streams. In this, they projected conventional image processing approaches to extract high-level a feature from a video stream. These features are altered into natural language text descriptions by means of the context-free grammar.

Premraj et al. [12] present a scheme to automatically create natural language explanations from input images that exploits together statistics collected from parsing huge quantities of text information and recognition algorithms from computer visualization.

Laptev et al. [13] tended to an acknowledgment of natural human exercises in differing and real-time video streams. This animating however vital subject has for the most part been disregarded in the past because of numerous issues one of which is the lack 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 evaluate elective approaches 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 problem of action classification in a video. Authors introduce a novel approach for video procedure that expands upon and extends a few late thoughts with space-time pyramids, neighborhood space-time highlights, and multichannel non-straight SVMs.

Laptev et al. [14] proposed a model for semantic justification of occasions, similar to weddings or b-ball games. The framework comprises event taxonomy, applied as a faceted classification, and an event partonomy, practical using the ABC ontology. Lee et al. [15] propose a high-level image illustration, called as the Object

Bank in which an image is showed as a scale-invariant map of enormous pre-trained common object locators, oblivious in regard to the testing dataset or visual task.

Li et al. [16] present a modest yet active technique to automatically compose image descriptions expected 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. Lin et al. [17] present the Google Books Ngram Corpus that illustrates how routinely words and expressions were used over a time of five centuries, in eight dialects. This technique 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.

Tanvi and Mooney [18] present combination of standard object recognition, activity classification, and text mining to study effective activity recognizers deprived of perfect labeling training videos. They create cluster verbs used to describe 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.

Packer et al. [19] presented a system that is able to recognize difficult, fine-grained human actions with the management of objects in truthful action sequences. Reddy et al. [20] 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. Wang et al. [21] 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.

Yang et al. [22] 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.

Yao et al. [23] give object and human position as the context of each other in different Human Object Communication (HOI) activity classes. They develop a random field model that uses a construction learning technique to learn significant connectivity patterns among objects and human body parts. Patil and Kagalkar [24] presented a method consists of two main modules such as image-to-text and text-to-speech using edge detection and image segmentation. An image-to-text module generates text descriptions in natural language based an understanding of the

image. A text-to-speech module converts natural language into speech synthesis.

Patil and Kagalkar [12] introduced image to text conversion need for blind people and system overview of an image to text and speech conversion system. Edge detection plays an important role in this system where the canny edge detection algorithm is used to detect objects from images. Object recognition is done based on color, size, texture, and shape of the object.

3. SYSTEM OVERVIEW

The proposed method can be viewed as natural language descriptions for visual content. First, an outline of the method as shown in Fig.1. The video splits into frames at one second intermissions. For each frame of input video images, the body and skin regions of a human are extracted by calculating the difference of colors between input and background images pixel by pixel. The positions of the head and the hands are found by perspective transformation. 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 every 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. To accomplish the effective image processing systems are utilized, for example, outline differencing based tracking, edge recognition to area shapes in videos. Examining shapes of the regions of a human and an object appeared on difference images, it can be verified whether the human being and put the object, or pick up and take it out.

The system presents a methodology for generating Natural language descriptions of long length videos by identifying the object and actions for descriptive videos. The proposed system consists of two major modules training and testing.

3.1 TRAINING MODULE

The training section is used to train videos and stored in the database with its features, objects and activity description which require for testing. Firstly, the video is split into images or frames since a video is nothing but a set of images. Training is performed on long length videos. After that, every Image is processed by filtering technique (noise removal, edge detection or shape detection) and applying Scale-invariant feature transform (or SIFT) feature extraction algorithm. SIFT algorithm is used to extract scale, orientation, and description of the image pixel. It takes a gray image as input. The input is a gray-level image. The output is a list of 2D points on the image each associated to a vector of low-level descriptors. These points are said key points and their descriptors are invariant by rescaling, in-plane rotating, and noise addition and in some cases by changes of the illuminant. SIFT describes an image or a portion of it by interest points (corners) whose detection requires a multi-scale approach Object and activities of objects are used to create exact sentences which are then ranked for likelihood as well as grammaticality. Thus, in the training section, objects and activities of each image are inserted into the database.

3.2 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. A 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 to training videos to recognize text in the English language. The prosed system under goes following step to yield the desired results,

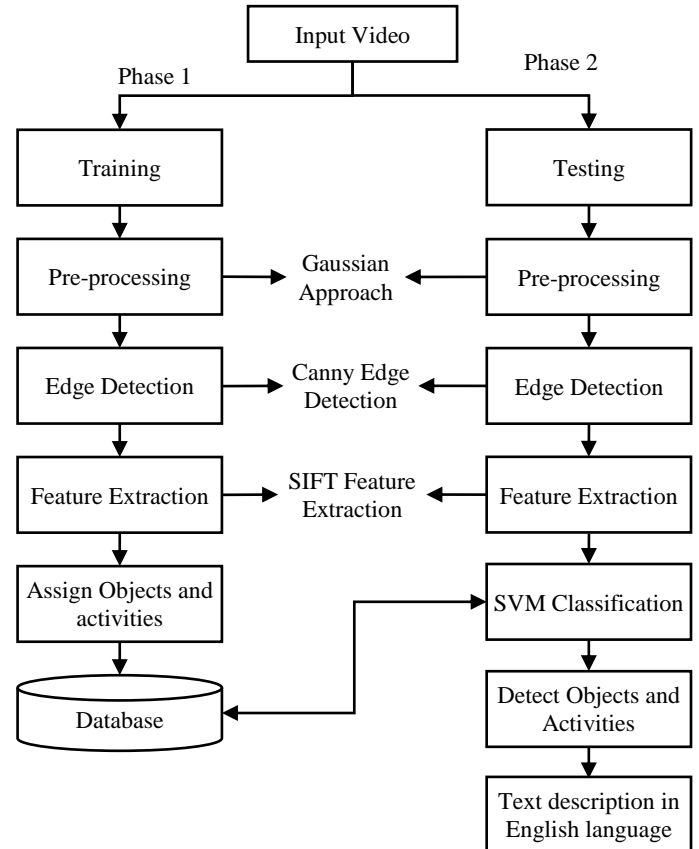


Fig.1. Proposed system overview

3.3 VIDEO ACQUISITION

It takes the video from the user as an input and performs translation of videos into its frames (multiple images). Then each frame undergoes preprocessing which is discussed in next section.

3.4 PREPROCESSING

This section gives description of preprocessing of frame. Preprocessing contains elimination of noise and blur and detects edge. Multiple frames from Video holds a 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.

A Gaussian filtering technique is used to eliminate blur from images and remove noise. Graphically Gaussian distribution can see as bell shape if mean is 0 and standard deviation of the distribution $\sigma = 1$.

The Canny edge discovery procedure is utilized to recognize the edges of objects present in pictures or frames. The Canny calculation essentially discovers edges where the gray scale 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.

3.5 FEATURE EXTRACTION

For any object, there are numerous elements, interesting points on the object that can be extracted to give a “feature” description of the object. This description can then be 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”. Each of these feature vectors is invariant to any scaling, resolution or interpretation of the image. To help the extraction of these elements the SIFT algorithm applies a 4-stage separating approach:

3.5.1 Scale-Space Extrema Detection:

In this process, there is a need to find a characteristic scale for the feature. Practically, maxima of Laplacian-of-Gaussian give the best notion of scale.

3.5.2 Keypoint Localization:

In scale space extrema get fewer points than pixels. And these points may include some bad points. To solve this Taylor series expansion is taken and minimize that to get the true location of extrema.

3.5.3 Orientation Assignment:

To set good points choose a region around each point using orientation. This uses a scale of point to choose correct image. This process is used to compute gradient magnitude and orientation.

3.5.4 Keypoint Descriptor:

Keypoint descriptor provides all descriptions of scale orientation, angle, and location of good points.

3.6 CLASSIFICATION

SVM classification is essentially a binary (two-class) classification technique, which must be modified to handle the multiclass tasks in real world situations. SVM classification uses features of image to classify the object and will be stored in data base. Generalization is the key idea behind classification. A classifier works well not only on the training samples, as well as previously unseen samples. From computational learning theory, SVM is based on the minimization principle of structural risk. It can learn the dimensionality of the feature space independently. This makes SVM to create general hypothesis from many features, which is the solution for text classification. Video usually includes large number of features. SVM works well with lots of features with ability to generalize well in high dimensional feature spaces, it makes the application easier for text categorization.

3.7 TEXT GENERATION

In this section, the important objects of the image are analyzed, then the matching and comparing of objects in the images with database are carried out. The related text for the recognized objects is generated.

4. ALGORITHM

An overview of the system is presented in Fig.1. For an input video sample, the following steps are followed:

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

Step 1. Convert Video into image frames.

Step 2. Apply Gaussian Filter for noise and blur elimination. The Standard deviation of the Gaussian function plays an important role in its behavior.

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

The Gaussian function is used in numerous research areas:

- It defines a probability distribution for noise or data.
- It is a smoothing operator.
- It is used in mathematics.

Gaussian function is never equal to zero. It is a symmetric function. The Gaussian filter is a non-uniform low pass filter.

Step 3. Apply Image edge detection algorithm is proposed based on morphology, canny edge detector.

The algorithm runs in 5 separate steps:

- *Smoothing:* Blurring of the image to remove noise.
- *Finding gradients:* The edges should be marked where the gradients of the image have large magnitudes.
- *Non-maximum suppression:* Only local maxima should be marked as edges.
- *Double thresholding:* Potential edges are determined by thresholding.
- *Edge tracking by hysteresis:* Final edges are determined by suppressing all edges that do not connect to a very certain (strong) edge.

Step 4. Apply SIFT Descriptor and extract image features.




Step 5. Apply SVM classification to identify the type of object and will be mapped with database to generate equivalent text description in English.

Step 6. The generated equivalent text description will be verified with grammatical checker to give grammatical correct text description of corresponding input video sample in English.

5. RESULTS AND DISCUSSION

To evaluate the text extraction process from videos, at least, 100 videos are used to train and store in database.

Table.1. Proposed System Result Description

Video Samples	Description			
	Actual Description	Expected Description	Processing Time	Video Size
	Person cooking roti. There is a rack fill of roti. A person sitting near shop. Women and a man selling sandwiches and food. A man driving a cycle. A man driving a scooter. A man pulling cart. One person is sited on seat of cycle.	In this video, there is a person. Person cooking roti. There is a rack fill of roti. There is person sited near shop. There are women and man selling sandwiches and food. There is man driving a cycle. There is a man pulling cart. Cart is full of luggage. One person is sited on the seat of the cycle.	75 Ms	30 Kb
	Kids are playing in water. Water tank is full of water. Water is falling from water tank. Two children sleeping. Some children enjoying water fall.	Kids are playing in water. Water tank is full of water. Water is falling from water tank. Two children sleeping on ground. Some children enjoying water fall. One kid is playing with play device. There is water park.	90 Ms	40 Kb
	A girl buying garments. Running auto. Cloths are there in road side shop. Car is passing near girl. Another car near auto. A shop for shoes and sandals. Two girls are buying sandals.	A girl is buying garments from a person. There is an auto on the road. Car is passing near girl. There is another car near auto on the road. There is shop for shoes and sandals. Two girls are buying sandals.	104 Ms	50 Kb

In the presented scenario, the testing video is trained before so that the presented tests previously include some video description to get the best result. The dataset is made of description of images that are used as training.

Some likely results are predicted using any long length of videos as shown in the table given below. Video is taken near about 40 seconds.

In above Table.1, video 1 is of market area. This video is capture for street food market area. For experiments, we have taken 30 second video having size 30kb. In this video one person is cooking food in his shop. Another person is serving food and buying it to his customer. The video contains so many objects like roasting pan having oil on gas; ladle used to cook food, there is cart on which some food is placed on some pans. There is a cycle rickshaw driving a person in front of cart. There are approximate 20 objects in video. This video 1 takes 75ms for processing.

The video 2 is of waterpark in which many children are playing different games. Water tank is full of water. Water is falling from water tank. Two children are sleeping. Some children enjoying water fall. Children are enjoying the water. This video duration taken is 45 seconds which is 40kb in size. Processing time require for this video is 90ms.

Finally, video 3 is used for experiments in which area of Indian market is shown. A girl is buying some neckless from a person who is selling on road. There are some shops behind that person. There is a car passing on road, a rickshaw standing on road. One person is selling some sweet on road. She is buying some footwear from a shop where two girls are buying for them and a guy is selling that footwear's. This video duration taken is 50 seconds which is 50kb in size. This is very complex video having numerous objects.

In Table.1 show the system tested for depending approach and per depending approach it gives 90% result of testing videos. This result is computed by using number of objects present in the video. System firstly train videos by inserting objects, activities and description into the database and tested some of them videos to evaluate result. The result consists of

1. Processing time
2. Video size
3. Grammatically correct sentences

For video sample 1 processing time require to extract grammatically correct sentences is 75 ms having size 30 kb. As per analysis, the output result sentences are good for this video thus table shows value 4 for grammatical correct sentences. Secondly, for video sample 2, which is video for waterpark, having size 40 kb, take time for generation of sentences is 90 ms. As shown in table2 , sample 2 gives output result 3, which means from analysis it gives ok result for grammatical correct sentences. Which means it generates some grammatical correct description for this video. Video sample 3 is very complex having numerous objects. Video sample 3 is capture for market area of India. There for the result sentences generated for this video are worst.

Table.2. Results summary of proposed system

Video samples	Processing Time (ms)	Video size	Grammatical correct sentences
Video sample 1	75	30 Kb	4
Video sample 2	90	40 Kb	3
Video sample 3	104	50 Kb	1

In Fig.2 shows a time require for processing video depend on the size of the video is shown. Time in milliseconds is required

for Gaussian filtering, Canny Edge detection and SIFT feature extraction. The graph shows that SIFT feature extraction requires more time as compare to Gaussian and canny process since it uses an internal Gaussian technique to process. As size of video increase, the time requires for processing that video gets increases. As shown in graph if video size is 20 seconds then the time require for processing of Gaussian technique, Canny edge detection and SIFT feature extraction is 55, 54 and 55 milliseconds respectively. Similarly, time require for Gaussian, Canny and SIFT process are computed for multiple videos of different size.

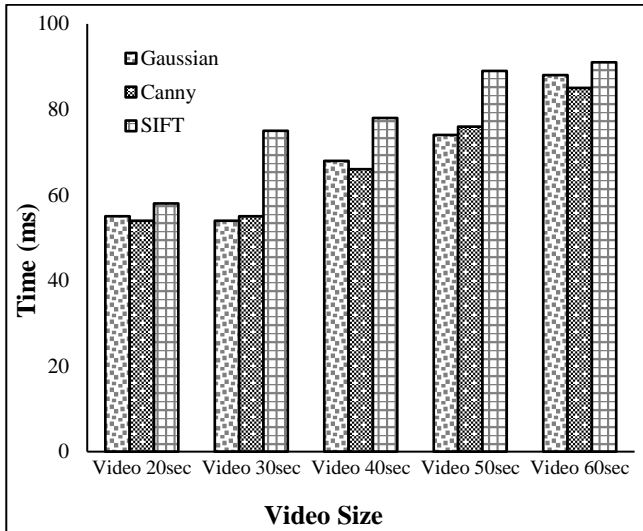


Fig.2. Time require for Gaussian, Canny, SIFT process

Following Table.3 shows the readings of graph represented above in Fig.2. For experiments, complex video are used having multiple objects with size in seconds. For example, for experiments video of 20 seconds, 30 seconds, 40 seconds, 50 seconds and 60 seconds are used to analyze results.

Table.3. Processing time for Gaussian, Canny and SIFT

Video Size (seconds)	Gaussian Processing Time (ms)	Canny Processing Time (ms)	SIFT Processing Time (ms)
20	55	54	55
30	54	55	75
40	68	66	78
50	74	76	89
60	88	85	91

6. CONCLUSION

This paper has introduced natural language descriptions of long videos by using SVM classification. The process uses object detection, text mining, activity recognition and feature extraction. Each video splits into frames at one-second interval, and the filtering, shape detection techniques are applied on every frame. Features are mined using SIFT algorithm and these features are used for comparison of testing with the training video. Future work includes methods to produce a 100% result, with an

improved processing time and produce grammatically correct sentences.

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 Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 12, pp. 2891-2903, 2013.
- [2] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko and S. Guadarrama, "Generating Natural-Language Video Descriptions using Text-Mined Knowledge", *Proceedings of 27th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 541-547, 2013.
- [3] Andrei Barbu et al., "Video in Sentences Out", *Proceedings of 28th Conference on Uncertainty in Artificial Intelligence*, pp. 102-112, 2012.
- [4] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal and Bernt Schiele, "Translating Video Content to Natural Language Descriptions", *IEEE International Conference on Computer Vision*, pp. 433-440, 2013
- [5] S. Gupta and R.J. Mooney, "Using Closed Captions as Supervision for Video Activity Recognition", *Proceedings of 24th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 1083-1088, 2010.
- [6] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A Library for Support Vector Machines", *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 1-27, 2011.
- [7] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning, "Generating Typed Dependency Parses from Phrase Structure Parses", *Proceedings of the International Conference on Language Resources and Evaluation*, Vol. 6, pp. 449-454, 2006.
- [8] Duo Ding et al., "Beyond Audio and Video Retrieval: Towards Multimedia Summarization", *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pp. 1-8, 2012.
- [9] Ali Farhadi and Mohsen Hejrati et al., "Every Picture Tells A Story: Generating Sentences from Images", *Proceedings of European Conference on Computer Vision*, pp. 15-29, 2010.
- [10] P. Felzenszwalb, D. McAllester and D. Ramanan, "A Discriminatively Trained, Multiscale, Deformable Part Model", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [11] Muhammad Usman Ghani Khan and Yoshihiko Gotoh, "Describing Video Contents in Natural Language", *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pp. 27-35, 2012.
- [12] 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*, Vol. 118, No. 3, pp. 14-19, 2015.

- [13] Ivan Laptev and Patrick Perez, "Retrieving Actions in Movies", *Proceedings of the 11th IEEE International Conference on Computer Vision*, pp. 1-8, 2007.
- [14] Ivan Laptev, Marcin Marszalek, Cordelia Schmid and Benjamin Rozenfeld, "Learning Realistic Human Actions from Movies", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [15] Mun Wai Lee, Asaad Hakeem, Niels Haering and Song-Chun Zhu, "Save: A Framework for Semantic Annotation of Visual Events", *Proceedings of IEEE Computer Vision and Pattern Recognition Workshops*, pp. 1-8, 2008.
- [16] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg and Yejin Choi, "Composing Simple Image Descriptions Using Web-Scale N-Grams", *Proceedings of 15th Conference on Computational Natural Language Learning Association for Computational Linguistics*, pp. 220-228, 2011.
- [17] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman and Slav Petrov, "Syntactic Annotations for the Google Books Ngram Corpus", *Proceedings of 50th Annual Meeting of the Association for Computational Linguistics System Demonstrations*, pp. 169-174, 2012.
- [18] Tanvi S. Motwani and Raymond J. Mooney, "Improving Video Activity Recognition using Object Recognition and Text Mining", *Proceedings of 20th European Conference on Artificial Intelligence*, pp. 600-605, 2012.
- [19] Ben Packer, Kate Saenko and Daphne Koller, "A Combined Pose, Object, and Feature Model for Action Understanding", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1378-1385, 2012.
- [20] Kishore K. Reddy and Mubarak Shah, "Recognizing 50 Human Action Categories of Web Videos", *Machine Vision and Applications*, Vol. 24, No. 5, pp. 971-981, 2013.
- [21] Heng Wang, Alexander Klaser, Cordelia Schmid and Cheng-Lin Liu, "Action Recognition by Dense Trajectories", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3169-3176, 2011.
- [22] Yezhou Yang, Ching Lik Teo, Hal III Daum and Yiannis Aloimonos, "Corpus-Guided Sentence Generation of Natural Images", *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 444-454, 2011.
- [23] Bangpeng Yao and Li Fei-Fei, "Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17-24, 2010.
- [24] Mrunmayee 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 Science and Research*, Vol. 3, No. 11, pp. 2164-2167, 2014.