IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

A Survey on Video Activity Recognition using Text Mining

Vishakha Wankhede[1], Ramesh M. Kagalkar[2]

M.E Student of Computer Engg Dept, Dr. D.Y. Patil School of Engg and Technology, Pune[1]

Research Scholar and Asst. Professor, Computer Engg Dept, Dr. D Y Patil School of Engg and Technology, Pune[2]

Abstract: Normal language, whether spoken, written, or typed, makes up much of human communication. A colossal quantity of this language describes the visible world either straight around us or in snapshots and video. This paper reviews a system to mechanically generate ordinary language descriptions from the video and a procedure that generates natural language descriptions for video. The framework is divided into two sections known as training and testing part. The training part is used to train the video with its description like pursuits of objects in that video. The testing part is used to experiment the video and retrieve the output as description of video evaluating videos stored into database. Combining Natural-language processing (NLP) with computer vision to generate English descriptions of visual information is an important area of active research.

Keywords: Natural-language processing (NLP), computer vision, video evaluation, language descriptions, Video processing.

I. INTRODUCTION

Computer vision has advanced to sense people, classify their actions, or to make a distinction between a large number of objects and specify their attributes. The output is often semantic representation encoding activities and objects categories. [6] While such representations can be well processed by automated systems, the natural way to communicate this information with humans is natural language. [1] Thus, this work addresses the problem of generating textual descriptions for videos. This task has a wide range of application in the domain of human-computer/robot interaction, generate summary descriptions of (web) videos, and automating movie descriptions for visually impaired people. Furthermore, being able to convert visual content to language is a central step in accepting the link between visual and linguistic information which are the richest interaction modalities available to humans.

Generating natural language descriptions of visual content is an intriguing task, but requires combining the fundamental research problems of visual recognition and natural language generation (NLG). While for descriptions of images, recent approach has planned to statistically copy the conversion from images to text [5,16, 18], most approaches for video description use rules and templates to generated video descriptions [14, 9, 2, 21, 3]. A methodology for the conversion from video to language descriptions in a two-step approach is discussed. In the first step system presents an intermediate SR uses a probabilistic model, following ideas used to generate video descriptions [5, 15]. Then, given the SR, it represents NLG as a modify problem, that is translating the SRs to natural language descriptions. In contrast to related work on video description, it contains the SR as well as the language descriptions from an aligned parallel body containing videos, semantic annotations and textual descriptions. This approach is compared to related work and baselines using no intermediate SR and/or language model. Second, the right level of verbalization need not to be defined manually. Instead, a study from a parallel training corpus the most relevant information to verbalize and how to verbalize it [5]. For this the methods from statistical machine translation are employed [20-28] [14].

(a) The correct ordering of words and phrases, referred to as surface realization in NLG is to be implemented.
(b)SR should be realized in language.
(c) The proper correspondence between semantic concepts and verbalization is to be mentioned, i.e. there is no need to define how semantic concepts are realized [17] [29,30,31].

In section II, literature survey of video content analysis is mentioned. III depicts the proposed framework review. And finally, Section IV concludes the paper.

II. LITERATURE SURVEY

In [1] the authors present a system to repeatedly generate natural language text from images. This system consists of two parts. The first part, content planning, smooth the output of computer vision-based detection and respect algorithms with statistics mine from large pools of visually graphic text to determine the best comfortable words to utilize to describe an image. The second step, exterior realization, chooses words to construct natural language sentence based on the predict content and general statistics from natural language. Girish Kulkarni, VisruthPremraj,

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Vicente Ordonez present multiple approaches for the surface realization step and evaluate each using regular measures of similarity to human generated situation descriptions. Authors also collect enforced choice human evaluation between descriptions from the proposed making system and descriptions from rival approaches. The proposed organism is very effective at produce relevant sentences for images. It also generates descriptions that are markedly more true to the definite image content than previous work.

In [3] the system produces sentential descriptions of video: who did what to whom, and where and how they did it. Action class is rendered as a verb, participant objects as noun phrases, properties of those objects as adjectival modiers in those noun phrases, spatial relations between those participants as prepositional phrases, and characteristics of the event as prepositional-phrase adjuncts and adverbial modifiers. In [2] system presents there have been a number of attempts at create 'video textures', that is, synthesize new (potentially infinitely extended) video clips based on existing ones. One method for accomplish this is to change each outline of the record into an Eigen space using major Components Analysis so that the creative sequence can be view as a mark during a low-dimensional space. Neill Campbell, Colin Dalton, David Gibson, David Oziem, Barry Thomas says that a new series can be generated by moving from side to side this space and creating 'similar' signatures. These signatures may be derived using an auto-regressive process (ARP). Such an ARP assume that the autograph has Gaussian statistics. For many sequences this assumption is valid, however, some sequences are strongly non-linearly correlated, in which case their statistical properties are non-Gaussian. These two methods by which such nonlinearities may be overcome. The first is by modeling the non-linearity automatically using a spline, and the jiffy using a mutual facade model. New video sequence created using these approaches have images never present in the original string and appear extremely convincing.

In [4] the authors Small particle aggregators non-specifically inhibit multiple disparate proteins, picture them therapeutically useless. They frequently emerge as false hits and thus need to be eliminate in high-throughput screening campaign. Computational methods have been explored for identifying aggregators, which have not been tested in screening large compound libraries. hanbingrao, zerong li, xiangyuan li, xiaohuama,choongyongung, hu li, xianghuiliu, yuzongchen used 1319 aggregators and 128,325 non-aggregators to expand a support vector equipment (SVM) aggregator identification model, which was veteran by four methods. The first is fivefold cross-validation, which show comparable aggregator and significantly improved non-aggregator identification rates against earlier studies. The second is the free examination of 17 aggregators discovered separately from the training aggregators, 71% of which were correctly known. The third is retrospective viewing of 13M PUBCHEM and 168K MDDR compounds, which predicted 97.9% and 98.7% of the PUBCHEM and MDDR compounds as non-aggregators. The fourth is retrospective screening of 5527 MDDR compound similar to the known aggregators, 1.14% of hich were predict as aggregators. SVM show somewhat better overall routine against two other machine knowledge methods based on fivefold cross-validation studies of the same settings. Molecular features of aggregation, extracted by a feature selection method, are consistent with published profiles. SVM showed substantial capability in identify aggregators from large libraries at low false-hit rates.

In [5] the authorsthe nomenclature software describe in this paper is premeditated to assist students in becoming capable at using the classification rules of organic chemistry. Features of the software include a graphics habit that converts structural formulas into drawing, the ability to decide correct IUPAC names for depict molecules, and the capacity to spot errors in names that are effort by the user. This paper describes the current stage of progress of the LISP-based software and discusses its success in dealing with alkanes, alkenes, alkynes, and associated halides.

In [6] the authors Given the deluge of compact disk content that is becoming on hand over the Internet, it is ever more vital to be able to successfully examine and arrange these large stores of in order in ways that go beyond browsing or joint filtering. In this paper, appraisal of previous work on audio and video processing, and describe the task of topic slanting multimedia summarization (TOMS) using natural language generation (NLG): given a set of routinely extract features from a video, a TOMS system will robotically generate a paragraph of natural language, which summarize the essential information in a video belonging to a certain topic, and for example provides explanations for why a video was matched and retrieve. Possible features include visual semantic concepts, objects, and actions, environmental sounds, and transcript from repeated speech recognition (ASR).

Florian Metze, Duo Ding, Ehsan Younessian, Alexander Hauptmann see this as a first step towards systems that will be able to discriminate visually similar, but semantically different videos, compare two videos and provide textual output or summarize a large digit of videos at once. In this paper, an approach of solving the TOMS dilemma is presented. Various features of visual concept features, green sounds and ASR text features from a given video, and develop a template-based NLG system to produce a textual recounting based on the extracted features. Experimental designs are presented for ad infinitum evaluating and improving TOMS systems, and present results of a pilot evaluation of our original system.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Table 1 Shows observations and comparison of technique used by different research groups working on text description of video content

Sr. No.	Citation	Work carried out so far	Issues
1	[1]	Present a system to automatically generate natural language descriptions from images.	The first type of image description approach utilizes existing text to describe query images.
2	[2]	Animators are interested in this, since many of the background shots in computer-generated movies are difficult and time-consuming to produce.	There have been several attempts at creating 'video textures', that is, synthesizing new video clips based on existing ones.
3	[4]	Small molecule aggregators non-specifically inhibit multiple unrelated proteins, rendering them therapeutically useless.	The possible inclusion of some yet-to-be-discovered aggregators in the ''non-aggregator'' class, which may affect the capability of SVM for identifying novel aggregators.
4	[5]	The nomenclature software described in this paper is designed to assist students in becoming proficient at using the nomenclature rules of organic chemistry.	IUPAC names for depicted molecules, and the capacity to identify errors in names that are input by the user.
5	[6]	Previous work on audio and video processing is reviewed, and define the task of top oriented multimedia summarization (TOMS) using natural language generation (NLG): given a set of automatically mined features from a video, a TOMS system will automatically generate a paragraph of natural language, which summarizes the significant information in a video belonging to a certain topic, and provides explanations for why a video was matched and retrieved.	The general concepts like "Apartments" or "Primate", which have a higher probability in general, are ranked at top.
6	[7]	Humans can prepare concise descriptions of pictures, focusing on what they find important.	Out of vocabulary words pose serious difficulties, methods are used to distributional semantics to cope with these issues.
7	[8]	A discriminatively trained, multi scale, deformable part model for object detection. Our system achieves a two-fold improvement in average precision over the best performance	For each positive bounding box in the training data, and apply the existing detector at all positions and scales with at least a 50% overlap with the given bounding box.
8	[9]	In order to provide natural language descriptions for visual content, this paper combines two important ingredients	SMT is a mature field with existing approaches achieving respectable results across many language pairs, see e.g. [17] for a review and tutorial.
9	[11]	Address recognition and localization of human actions in realistic scenarios.	A new annotated human action dataset and use it to evaluate several existing methods.
10	[12]	The aim of this paper is to address recognition of natural human actions in diverse and realistic video settings.	Compared to the existing approaches, this method shows significantly better performance, outperforming the state-of-the art in the same setup.
11	[13]	Propose a framework that performs automatic semantic annotation of visual events (SAVE). This is technology which enables content-based video annotation, query and retrieval with applications in internet video search and video data mining.	Video search tools rely mainly on user annotated tags, captions, and surrounding text to retrieve video based on broad categories.
12	[14]	A unified framework for image descriptors based on quantized joint distribution of filter bank responses and evaluates the significance of filter	The presented framework contributes to understanding and comparison of existing texture descriptors but it can be

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

		bank and vector quantized selection.	utilized for more systematic development of new, even better performing methods.
13	[15]	Describe some theoretical and practical issues raised during the construction of the Basque Dependency Treebank (BDT): the syntactic annotation of EPEC (Reference Corpus for the Processing of Basque)	The theoretical and practical issues raised during construction of the BDT, following the Dependency Grammar theory
14	[17]	Present a system that is able to recognize complex, fine-grained human actions involving the manipulation of objects in realistic action sequences. Our method takes advantage.	The z values are accurate and are not sensitive to issues such as lighting and shadows that present major challenges for vision algorithms when dealing with images from traditional cameras.
15	[18]	Challenges like camera motion, different viewpoints, large interclass variations, cluttered background, occlusions, bad illumination conditions, and poor quality of web videos cause the majority of the state-of-the-art action recognition approaches to fail.	A framework that can address issues with real-life action recognition datasets.
16	[19]	Web videos cause most the state-of-the-art action recognition methods to fail. Also, an increased number of categories and the presence of actions with high confusion add to the challenges.	A framework that can address issues with real-life action recognition datasets.

As per the review of above paper workwe found that some major issues are still untouched, so this paper will try to solve the issues and further define in the proposed system which is discussed in next section.

III. PROPOSED SYSTEM METHODOLOGY

The proposed system introduces a holistic data-driven methodology for generating descriptions of long videos by recognizing the videos [32-33].

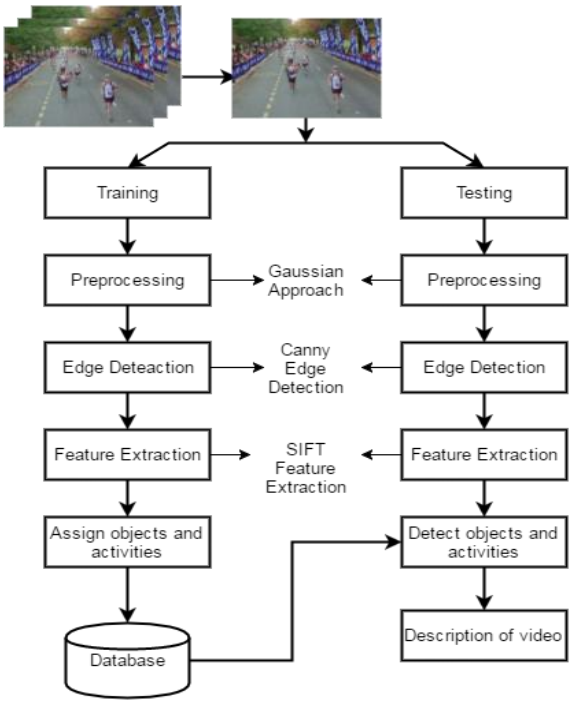

Fig.1 System Architecture

The proposed system consists of two major modules training and testing as shown in figure 1.

1. Training Module

The training section is use to train videos and stored on database with its features, objects and activities description which need for video testing. Firstly, the video is split into images or frames since video is nothing but a set of images. Training is performed on short videos because frames of long video are more. If there are number of images is more than time require to process one video will more. After that every Image is processed by filtering (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 plausibility as well as grammaticality [34]. This section is handled by admin who is accountable for data training.

2. Testing Module

This module test video learner and gets result if at slightest one video is trained. In this phase, video is processed and split into frames and these frames are further processed by applying filtering algorithm to remove noise from images. Gaussian filtering technique is used to filter image [35]. After removal of noise, the description of images is extracted to detect objects. These features are comparing with training videos to recognize text [36].

IV. CONCLUSION

This Paper has introduced natural language descriptions of long videos by SVM classification for complex videos containing more than 10 objects. The process uses object detection, text mining, activity recognition and feature

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

extraction. Each video splits into frames at one-second intervals 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.

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] 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.
[4] Chang, C., and Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (TIST) 2(3):27, 2011.
[5] 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.
[6] 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.
[7] 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.
[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] 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.
[10] 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.
[11] Laptev, I., and Perez, P., "Retrieving actions in movies", In Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV), 1–8, 2007.
[12] 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.
[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] 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.
[15] 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.
[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] Reddy, K., and Shah, M., "Recognizing 50 human action categories of web video", Machine Vision and Applications 1–11, 2012.
[19] 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.
[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] 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.
[22] MrunmayeePatil 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.
[23] 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.
[24] Vandana D. Edke and Ramesh M. Kagalkar, "Video Object Description of Short Videos in Hindi Text Language", International Journal of Computational Intelligence Research, Volume 12, Number 2 (2016), pp. 103-116 © Research India Publications.
[25] Vandana D. Edke and Ramesh M. Kagalkar," Video Objects Description in Hindi Text Language ", International Journal of Managing Public Sector Information and Communication Technologies (IJMPICT) Vol. 7, No. 3, September 2016.
[26] Ramesh M. Kagalkar and Dr. S.V. Gumaste,"Review Paper: Detail Study for Sign Language Recognization Techniques" CiiT international journal of Digital Image Processing, Volume 8, No 3 (2016)
[27] Ramesh M. Kagalkar, Dr. Nagaraj H.N and Dr. S.V Gumaste, "A Novel Technical Approach for Implementing Static Hand Gesture Recognition", International Journal of Advanced Research in Computer and Communication Engineering, Voume l. 4, Issue 7, July 2015.
[28] Ramesh M. Kagalkar, Dr. Nagaraj H.N," New Methodology for Translation of Static Sign Symbol to Words in Kannada Language", International Journal of Computer Applications (ISSN: 0975 – 8887) Volume 121 – No.20, July 2015.
[29] Ramesh M. Kagalkar and S.V Gumaste, "Gradient Based Key Frame Extraction for Continuous Indian Sign Language Gesture Recognition and Sentence Formation in Kannada Language: A Comparative Study of Classifiers", International Journal of Computer Sciences and Engineering, Volume-04, Issue-09, Page No (1-11), Sep -2016, E-ISSN: 2347-2693.
[30] Ramesh M. Kagalkar and S.V Gumaste, "New Frame Work for Translation of Sign Language Action into Text Description in Kannada", CiiT international journal of Digital Image Processing, Vol 8, No 10 (2016)
[31] AmitkumarShinde and Ramesh M. Kagalkar, "Sign Language Recognition for Deaf Sign User", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ©IJRASET, Volume 2, Issue XII, December, ISSN: 2321-9653, 2014.
[32] Amit kumar and Ramesh Kagalkar, "Methodology for Translation of Sign Language into Textual Version in Marathi", CiiT, International Journal of Digital Image Processing, Volume 07, No.08, Aug 2015.
[33] AmitkumarShinde and Ramesh M. Kagalkar," Advanced Marathi Sign Language Recognition using Computer Vision", International Journal of Computer Applications, (ISSN:0975 – 8887), Volume 118, No. 13, May 2015.
[34] Rashmi. B. Hiremathand Ramesh. M. Kagalkar, "Methodology for Sign Language Video Interpretation in Hindi Text Language", International Journal of Innovative Research in Computer and Communication Engineering, Volume. 4, Issue 5, May 2016.
[35] Rashmi. B. Hiremath and Ramesh. M. Kagalkar," Sign Language Video Processing for Text Detection in Hindi Language", International Journal of Recent Contributions from Engineering, Science and IT, Volume 4, No 3, 2016.
[36] Rashmi. B. Hiremath and Ramesh. M. Kagalkar" A Methodology for Sign Language Video Analysis and Translation into Text in Hindi Language", CiiT International Journal of Fuzzy Systems, Volume 8, No 5, 2016.