Video2Text: Multi-instance & Multi-level Region Annotation

Yogeshwari Gowda
Department Of Electronics and Telecommunications
Mumbai, India

Shruti Bore
Department Of Electronics and Telecommunications
Mumbai, India

Mayur Nehete
Department Of Electronics and Telecommunications
Mumbai, India

Akula Venkat
Department Of Electronics and Telecommunications
Mumbai, India

Prof. Dr.Sinora Banker
Professor,
Department Of Electronics and Telecommunications
Mumbai, India

***Abstract*— Picture book is a highly aesthetic genre, in which words and pictures together tells a story. Meanings in picture books are inextricably constructed by art and text. Picture books are different from the traditional story books in which pictures are used to supplement a text. Story books could be understood without reference to them. Illustrations undoubtedly explain the story to the readers, but some storybooks can be understood without them. Real-world videos often have complex dynamics and methods, for generating open-domain video descriptions and they should allow both input (sequence of frames) and output (sequence of words) of variable length. To approach this problem, we propose a novel end-to-end sequence-to-sequence model to generate text for videos. For this we are using Recurrent Neural Networks (RNN), specifically LSTMs, which have demonstrated state-of-the-art performance in image to text generation. Our LSTM model is trained on video-sentence pairs and learns to associate a sequence of video model is trained on video-sentence pairs and learns to associate a sequence of video frames to a sequence of words to generate a description of the event in the video clip. Our model naturally is able to learn the temporal structure of the sequence of frames as well as the sequence model of the generated sentences, i.e. a language model.**

Keywords- Long Short Term Memory (LSTM),RNN (Recurrent Neural Network)

I. INTRODUCTION

Image to text annotation is a challenging research topic due to the requirements of the knowledge of both vision modality and natural language modality. The ultimate goal of image to text annotation is to generate natural language description for given image in a real-time , just like what we humans do. Moreover, the generated language should be capable of describing the objects and their relations in the image in a grammar error-free and fluent way.

Despite the difficulty of this task, there have been emerging efforts recently due to the success of introducing the deep neural networks to this field. Most existing work leverages the deep convolution neural networks (CNN) and the recurrent neural networks (RNN) in an encoding-decoding scheme. In these work, the input image is usually encoded by a fixed length of CNN feature vector, functioning as the rst time-step input to the RNN; the description is generated by Conditioning the output word at each time step on the input visual vector as well as the previous generated words.

We develop a video captioning system, which allows the users to 1) generate human-level natural language description of input video, 2) detect objects in the given video alongside caption generation, and 3) retrieve similar images and descriptions from a database. The core of the system is a combination of pre-trained deep convolution neural networks for object detection, and recurrent neural networks for caption generation.

II. FLOW OF WORKING

1. Take data set (images and videos both are relevant/similar/same)
2. Extract features from images (sift, gist)
3. Make clusters of feature (Bag of words)
4. Classify clusters (manually)

Take test video extract frames
1.Extract features from images (sift, gist)
2.Mach features from bag of words using nearest neighbor algorithm
3. List out maximum confidant classes as annotation

Text mining from annotations
1. Generate database relevant to test samples (one can opt online documents)
2. Build an N gram Language Model
3. Take tags extracted from videos
4. Use of NLP using N gram for auto sentence / correct sequence/ words generations

III. SYSTEM FLOW

1) First the input video has to be taken
2) In the process of removing unwanted or redundant data from video we use Key frame extraction i.e. most relevant data images are obtained (using histogram analysis).
3) Scale-invariant feature transform (or SIFT) is an algorithm in computer vision to detect and describe local features in

images. SIFT key points of objects are first extracted from a set of reference images and stored in a database. An object is recognized in a new image by individually comparing each feature from the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors.

The GIST descriptor has recently received increasing attention in the context of scene recognition. The idea is to develop a low dimensional representation of the scene, which does not require any form of segmentation.

4) The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model has also been used for computer vision. The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier.

5) As an alternative, the n-gram model can be used to store this spatial information within the text. Applying to the same example above, a bigram model will parse the text into following units and store the term frequency of each unit as before.

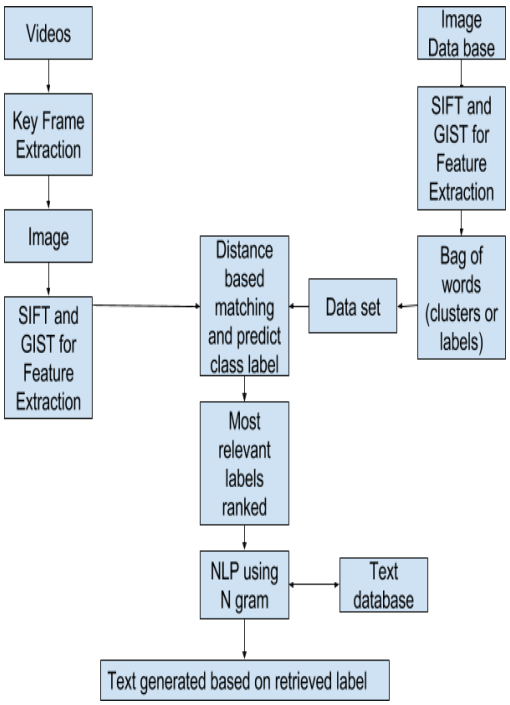

Figure 1. An overview of Video2Text.

IV. ALGORITHM

A. *GIST(Graphics and Intelligence Based Technology)*

GIST is a global descriptor. Observers

Steps:

1. Partition image into a 4x4 grid
2. Calculate histograms in multiple orientation channels
3. Calculate color/intensity histograms for each cell

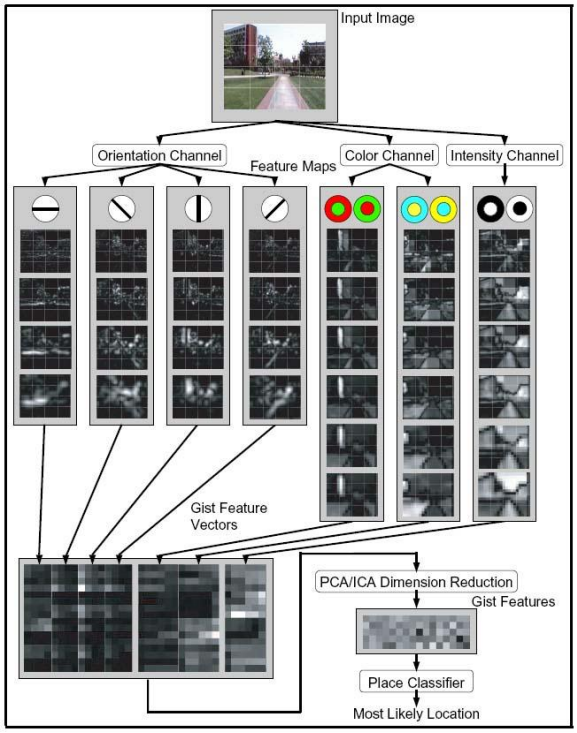

Figure 2. Use of GIST Descriptor

B. *SIFT (Scale-Invariant Feature Transform)*

It is a local descriptor.

To detect and describe local features in images.

Steps:

1. Find the blurred image of the closest scale
2. Rotate the region by the dominant orientation
3. Divide the region into a 4x4 grid
4. Create a histogram for each cell (8 angels)
5. Normalize the vector

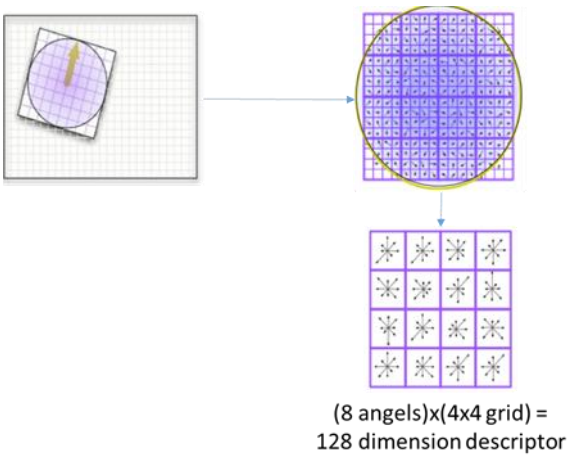

Figure 3. Use of SIFT Descriptor

C. *N-Gram Model*

NLP is a probabilistic model of a language that gives a probability that a string is a member of a language is more useful. It specify a correct probability distribution, the probability of all sentences in a language must sum to 1.

Here we are using this for speech reorganization.

Steps:

1. A language model also supports predicting the completion of a sentence.
2. Predictive text input systems can guess what you are typing and give choices on how to complete it.

V. DATABASE

Category	Number of Video
Mall	50
Airport	50
Parking	50
Traffic-Road	50
Total	200

VI. EXPERIMENTS AND RESULTS

We first iterated the code and removed bugs to implement our first conversion from image to text. The tested system is MATLAB. It becomes very difficult to compare the two image. So, instead we tried comparing the two different image with the same object. Our method evaluated give precision between the same object and convert into textual format. We designed our experiments to get performance. In the first experiment we took normal objects such as elephant, butterfly etc . We tried to keep the object same with different angle so as to get proper recognisition. In the final experiment we were able to get the accuracy for our output.

VII. PROBLEM DISCUSSION AND SOLUTIONS

The problem arises when the algorithm used was not able to get proper output. We tried using variour algorithm but atlast, GIST algorithm were able to get the exact matching of our object. The difficulty arises to test the train image and test images. It was necessary to link the path between the two test series. The coding was very tedious.

CONCLUSION

The project focuses on converting video data into corresponding text. We use key frame extraction for extracting important images. After that we use SIFT and GIST for feature extraction. These are compared with the data base and the text corresponding to the image is obtained.

VIII. FUTURE SCOPE

In the future, the project can be extended for classification of other modalities including text as well as video. The text modality deals with the detection of the text which is present on the screen i.e. indexing and searching is done on the basis of words which are found onscreen. The visual modality i.e. video modality deals with pattern matching and image mining performed on the frames extracted from the video. A combination of two or more modalities can also be used. This project is a Desktop application, in future it can be extended as a server application.

ACKNOWLEDGMENT

We would like to express our deepest appreciation to all those who provided us the possibility to complete this synopsis. Our first and foremost thanks goes to Atharva College of engineering. A special gratitude we give to our final year project guide, Dr. Sinora S. Banker, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project with all required equipment and the necessary materials to fulfill the requirements regarding "*Video2Text:Multi-Instance & Multi-Region Annotation*". A special thanks goes to HOD Prof. Jyoti Kolap & to the coordinator of the project, Prof. Prajakta Pawar who have invested her full effort in guiding the team in achieving the goal.

REFERENCES

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks a for visual recognition and description. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation.
[3] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding long-short term memory for image caption generation.arXiv preprint arXiv: 1509.04942, 2015.
[4] Applications and Algorithms in CV- tutorial9
[5] https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwi__rX4rp7RAhWHNo8KHVQoBjMQFggmMAI&url=http%3A%2F%2Fdb.cs.berkeley.edu%2Fpapers%2Fvldb95-gist.pdf&usg=AFQjCNHPg_dvcKWQy_VQzYIv29ElOloltw&sig2=lpzgrEgy8CG5FiLwgKkbZQ
[6] https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=7&cad=rja&uact=8&ved=0ahUKEwjtx5nXr57RAhVFtY8KHfpxB5QQFgg8MAY&url=http%3A%2F%2Fwww.wisdom.weizmann.ac.il%2F~vision%2Fcourses%2F2004_2%2Ffiles%2Fgist%2FGistPresentation.ppt&usg=AFQjCNE654HWF7g6NHTCCwSeoEaKqkfXSQ&sig2=qbKSo_iX-ZxLDl34GM0h0Q
[7] https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&cad=rja&uact=8&ved=0ahUKEwjtx5nXr57RAhVFtY8KHfpxB5QQFghLMAg&url=https%3A%2F%2Flear.inrialpes.fr%2Fpubs%2F2009%2FDJSAS09%2Fgist_evaluation.pdf&usg=AFQjCNHOBoA86kJA1sGH68Z7RZgurK6ucA&sig2=7Xk3TdFgNcvnT0bEMLMmPg