Available online at: http://www.ijmtst.com/vol6issue04.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 06, Issue No: 04, April 2020

A Survey on Identification of Objects from Images Based on Spoken Words

Riddhi Surani[1] | Dr. Darshak G. Thakore[1] | Dr. Udesang K. Jaliya[1]

[1]Department of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Vallabh Vidyanagar, Gujarat, India

To Cite this Article

Riddhi Surani, Dr. Darshak G. Thakore and Dr. Udesang K. Jaliya, "A Survey on Identification of Objects from Images Based on Spoken Words", *International Journal for Modern Trends in Science and Technology*, Vol. 06, Issue 04, April 2020, pp.:287-292.

Article Info

Received on 25-March-2020, Revised on 15-April-2020, Accepted on 19-April-2020, Published on 25-April-2020.

ABSTRACT

Identifying objects from an image and establishing their relationship is a very critical task for understanding the image. There are several contributions provided in the past that identify objects from the image and provide a textual label to the objects. Nowadays, use of Deep Learning based methods came into the picture due to an increase in computational power. Recently, some approaches had been proposed to link spoken words with the objects in the image. This type of linking can help in searching for the image using voice commands. This is an application for both computer vision and natural language processing. In thispaper, some of the recent approaches in the field of object identification using speech data are discussed briefly. It also presents some popular dataset details and at last, brief discussion is provided on object detection methods and models.

KEYWORDS: *Object Detection, Speech Network, Convolutional Neural Network, Multi-Modal Learning*

I. INTRODUCTION

Identifying objects and their relationship from an image is an interesting and challenging problem. It will extract the objects from the given input image based on the words from the spoken caption.

Object Detection based on spoken words is considered as one of the complex AI problems as it involves both image and speech. So, it has a large research community from both Computer Vision and Natural Language Processing. It is still very trending area for research, despite of many years from its origin.

The purpose of the object identification from spoken words is to detect all the occurrences of objects by the classification and localization from a given image. It will detect the object from the image which has a different class like the person, vehicle, animals and many more[1]. Several contributions in the past are provided based on identifying objects from the image and gives a textual label to the objects. Instead of text, as in the earlier approaches, it uses the speech data for some applications that will not be able to use text data. This system refers to analyzing an input image and spoken captions, extracting objects and words, establishing relationships among the objects and words.

There is a wide range of applications for Object Detection using spoken words, which can be helpful in the day to day life. It can be used for searching image from the large collection of images and also to craft the video using this image data and audio data.

In Figure 1, a sample image and output is provided with a spoken caption. The process of Object Identification stipulates identifying the object from the input caption. It shows the individual words with an associated confidence for an input spoken caption.

Figure 1: Identified objects with the Spoken caption[2]

In this paper, the overall introduction of the research is presented and an important concept in a relatively simple manner is also provided briefly. The organization of the report results in five sections which includes a brief Introduction, an Object Identification system which provides approaches for Object Identification using spoken words. The third section is about a survey on the existing methods which includes several Object Detection methods with a brief description. The later sections present the dataset details which contains images with the audio captions and a summary about the presented paper.

II. OBJECTIDENTIFICATION SYSTEM

This system is also known as a multi-modal learning task where the modal has more than one different inputs and it provides the relationship between them[3].There are different approaches to achieving this problem. One of the approaches is Supervised Learning and other approach is Unsupervised Learning. In the Supervised Learning approach, it provides the pre-defined objects from an image and segmented words from the caption as an input. It provides the label to the detected objects from an image. In the Unsupervised Learning approach, it does not require any supervision. It directly performs on the image pixels and speech waveform rather than on the object labels. It introduces a Novel Neural Network[4].

Identification of objects using audio commands make use of Deep Learning techniques, thus the focus of this paper is on Deep Learning methods. Many of the researchers in the past had used a Deep Learning approach.

David Harwath and James Glass[2] proposed the model which takes an input image along with the corresponding spoken captions and find the correlation between them. In this approach, it presentedan Audio-Visual Alignment Model that makes use of pre-segmented inputs. The pre-segmentation step identifies the objects using Region Convolutional Neural Network (RCNN). Later, the extracted regions are provided to the Convolutional Neural Network (CNN) and the audio captions are provided to the Spectrogram Convolutional Neural Network. To merge the outputs from both CNN and Spectrogram CNN, an Embedding Alignment model is used. StochasticGradient Descent is implemented in order to optimized the cost function.

D.Harwath, A. Torralba, and J. Glass[4] proposed the model that finds words and phrases directly from the audio waveform. It uses the novel neural network for object identification. It takes image and audio as an input and extract features from both image and audio. It uses VGG16[23]and fliter bank for image and audio feature extraction respectively. The image and audio network will measure similarity score using dot product. The dot product of image and audio caption vector is train using stochastic gradient descent using an objective function which measures the similarity score of the matched image and audio caption pair and the mismatched image and audio caption pair. Flickr8k[5] and Flickr30k[6] is used as image data and Audio captions is collected via Amazon MechanicalTurk[25].

David Harwath and James Glass[7] proposed the model which jointly learns the semantics of the word-like units through visual associations. This model is able to detect the spoken word and associate it with that region in an image. This model is made up of two networks. One which have input as an image and another which have input as spectrogram. The image network is formed using the VGG16[23] layer network and the spectrogram is computed using the Log Mel filterbank with Hamming window and shift. The Multimodal is formed by joining the image network and an audio

network. It combines the output vector of the both network and perform the inner product of them. Stochastic Gradient Descent is used to train the neural network. To perform clustering of Audio-Visual caption grounding it uses the constrained brute force ranking scheme to compute all possible similarities between image and their caption. The K-Means clustering separately apply on the both collection of images embedding vectors and the collection of acoustic embedding vectors.

D.Harwath et al.[8] proposed the model that directly operates on image pixels and speech waveforms. It does not require any label, alignment, or segmentation. A paired image and audio are provided as an input. The author proposed an Audio-Visual MatchMap model. A VGG16 [23] model is used for feature extraction from image and performed Convolutional Neural Network on both image and audio. For computing the image-speech similarity, three functions were proposed namely Sum of Image and Sum of Audio (SISA), Maximum of Image and Sum of Audio (MISA), and Sum of Image and Max of Audio (SIMA). They used the ADE20k dataset[9]for image and collected 100,000 captions forthat.

A brief overview of some related approach due to which the idea for object identification using speech data developed. Those approaches use text transcription instead of audio and some of the also presented multi modal for image and text.

Chen Kong et al.[10] describes the problem of aligning text to the images. It parses the text from the caption and identifies the visual objects from the image and then aligns these two modalities. It is a multimodal learning task. The alignment for the text and image is provided using a Markov random field.

Andrej Karpathy et al.[11] presented the model of text-image alignment. In this paper, using Region Convolutional Neural Network (RCNN), it detects the visual objects from the image and they are mapped to fragment embedding space. Through this process, learning for image and word can be performed. It provides interpretable predictions for the image sentence retrieval task.It uses the two Objective functions to measure the image-sentence similarities reliable with their ground truth. The first function is Fragment Alignment Objective and the other is the Global Ranking Objective. Stochastic Gradient Descent is implemented in order to optimize the cost function.

Chetan Amritkar and Vaishali Jabade[12]proposed the model which generates caption for the images. Their model proposed two neural networks. First is the Convolution Neural Network which is used for feature extraction from the image and the second, Recurrent Neural Network for the sentence generation. To generate this model, they used Flickr8k[10] dataset.

III. SURVEY OF EXISTING METHODS

The identification of objects based on spoken words can be implemented by using various methods. It includes Object Detection, Feature Extraction for image and audio, and alignment of object and word. Object Detection is the combination of two tasks, Image Classification, and Object Localization [1]. Image classification includes assigning a class label to an image[1], where object localization includes the identification of one or more objects located in an image and drawing a bounding box around one or many objects in an image[1]. A Spectrogram Neural Network is used for processing the speech data for feature extraction.

The Figure-2 shows a general workflow of this system from the referred literature [3]. It takes input as image and audio and then applies the Convolutional Neural Network for both image and audio and then both the image and audio network perform the Match Map. At last, it computes the similarities for the image and audio. Three different functions are proposed to compute the similarities between the image and their spoken caption. As it takes image and audio as an input it is called a multimodal learning task. These methods and models will be discussedbriefly in thissection.

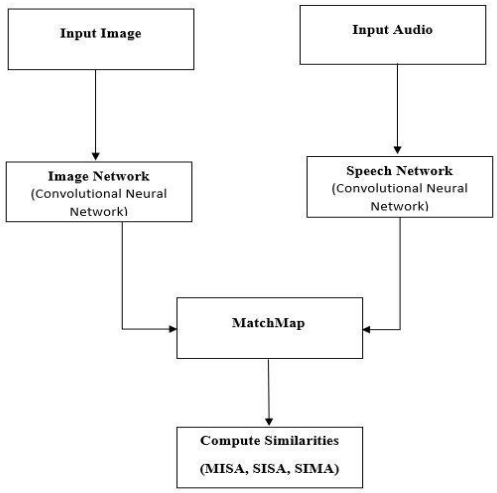

Figure 2: Existing workflow [3]

Detection. YOLO method attained a wide acceptance due to faster in nature compared to R-CNN methods and it can also achieve real-time object detection [13]. As this system includes speech data and for their feature extraction, a Spectrogram Convolutional Neural Network is used

A. Region Convolutional Neural Network

Region convolutional neural network [2] used to detect the region in the image which contains the objects. The RCNN used the selective search to extract the regions from the image. RCNN model contains the four steps. First, it takes the input as image and then it extracts the region proposal that are warped into a square and that region proposal fed into convolutional neural network[13]that produces the feature vector and in the last step,extracted feature fed into support vector machine to calculate detection scores over some set of classes for each region and gives the classify regions.

B. You Look Only Once[15]

You Look Only Once (YOLO) [15] used to detect the object in the image. YOLO model is faster than other models it is main advantage of this model. First step is to divide the input into S*S grid cells. Second step is to detect the object: center of an object comes into that grid cell then that grid cell is detecting the object. Third step is to predict the bounding box and confidence score. Confidence scores tell that how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts.

C. Convolutional Neural Network[16]

The Convolutional neural network [16] is used to do image classification, image recognition, object detection and many. CNN processes an input image and categorizes it into classes. An input image passes through several convolutional layers. There are four layers in a CNN: Convolutional Layer, Max Pooling Layer, Flattening and Fully Connected Layer. The first layer is the Convolutional Layer and which is used for extracting features from the input image. It establishes the relation between pixels of an input image and features detector. The feature is called the kernel or it is called a filter. It is just element-wise multiplication on these two matrices. Based on the type of kernel, Feature Map is generated as an output. The activation function that is used by CNN is ReLU, which is known as the Rectified Linear Unit. The reasonbehind applying the rectifier is to increase the non-linearity in the image. The second layer is the pooling layer. The layer reduces the number of parameters from an image. It contains three Pooling layers - Max pooling, Average pooling, and Min pooling. Max pooling takes the largest element and min pooling takes the smaller element from the feature map. The third layer is the flattening which takes the number row by row and puts them into one column and the last layer is Fully Connected layer in which flattened vector is fed into a neural network and which uses Soft Max or sigmoid activation function to classify the outputs for classes.

D. Spectrogram Convolutional Neural Network [2]

Spectrogram Convolutional Neural Network is used to do word classification on an audio caption. It is using a similar Convolutional Neural Network to model the spectro gram for each word in the audio caption. It is also containing the 4 layers Convolutional Layer, Max Pooling Layer, Flattening and Fully Connected Layer.

E. Embedding Alignment Model[2]

The embedding alignment model is used to align the image and audio model. The image and their corresponding caption pair have their corresponding object detection box and the word spectro gram.To achieve this task the transfer modelfrom[11] is used with objective function[17]. It takes the dot product between the two embedding vectors and then take sum over the words it gets similarity score.

F. Matchmap Model [8]

D.Harwath et al.[8] proposed match map model to compute the similarities between the image and speech pair wair wise. Here they present the three similarity score functions. The first function is SISA it gives the sum of all the images and sum of all the audio. As in reality, it is not completely expectable to match words within a caption to match all objects simultaneously within an image. So, the second function is MISA it gives the max of all images and sum of audio. It matches each caption with the most similar image and the third function is SIMA it gives the sum of all images and max of audio. Among all three functions, Max of all images and sum of audios (MISA) function is the best performing similarity function.

IV. DATASETS

Identification of objects based on spoken words contain images and their respective spoken captions. Some of the widely used datasets for identification of objects based on spoken words are Flicker8k[5], Flicker30k [6], MSCOCO [18], and ADE20K[9].

A. Flicker8k[5]

Flicker8k[5] contains 8000 images extracted from Flicker. It mainly contains natural images. Each image is annotated by five sentences and Flickr8k Audio Caption Corpus contains 40,000 spoken captions of 8,000 natural images which is collected via amazon mechanical Turk[25].

B. Ade20k[9]

The Ade20k[9] dataset contains 20,210 train images and 2000 test images. This dataset contains nearly 200,000 recordings which is collected via Amazon Mechanical Turk that describing the images[3]. It collects the one caption per image. The total number of image caption pair is 402,385[8].

C. Mscoco[18]

The Mscoco[18] image dataset contains natural photographs covering huge variety of types of scene. For the image, coco dataset is split in 3 different types of dataset that are 2014,2015 and 2017. In a 2014 there are 83k train images, 41k test images and 41k validation images. Each image paired with set of five caption. In a 2015 there are 81k test images and at last, in 2017 there are 118k train images, 5k validation images, 41k test images and 123k unlabeled images. The Mscoco audio corpus[19] of 2014 dataset which contain 616,767 spoken captions in which for train2014 it contains 414,113 and for val2014 it contains202,654.

Here, Table 1 provides the sample images and their text transcription of associated audio captions of Flicker8k[5], Ade20k[9] and Mscoco[18] datasets.

Table 1. Sample Dataset Images and Text Transcription of their Associated Spoken Captions[5],[9],[18],[19],[24]

Dataset Name	Sample	Text transcripts of their associated Spoken captions
Flickr8K [5]		1075716537_62105738b4.jpg#0 A child with a helmet on his head rides a biksse . 1075716537_62105738b4.jpg#1 A little boy rides a bike down a hill on a miniature dirt bike . 1075716537_62105738b4.jpg#2 A young boy in a helmet rides a bike on the road . 1075716537_62105738b4.jpg#3 The little boy rides his bicycle in a race . 1075716537_62105738b4.jpg#4 The young boy pedals quickly at a BMX race .
Ade20K [9][24]		"Young boy standing on a tire swing he's wearing a black and white striped shirt."
MSCOCO [18]		{"image_id":100012,"id": 243584,"caption":"The player in white is ready to caught the frisbee."} {"image_id":100012,"id": 210296,"caption":Oh . Two men in field catching a white frisbee."} {"image_id":100012,"id": 249257,"caption":"Two men go after a frisbee on a soccer field."} {image_id":100012,"id": 230903,"caption"An image of a kids playing with a frisbee."} {"image_id":100012,"id": 230288,"caption": "Two people on a field trying to catch a frisbee."}

V. SUMMARY

The identification of objects based on spoken words is a technique that identifies the objects from the given images. The object identification using

speech data is still a very fast- growing research field. It is the application of computer vision and natural language processing Every day, new works are published in this field. Our effort in this work is intended to give a brief conceptual overview of the field and how it works. We have summarized some important models and methods for the classification which is in the existing work. And, we have also taken a basic overview of some popular datasets for images and audios corpus such as Flicker8k[5], Flicker30k [6], MSCOCO [18], and ADE20K[9]and some remarkable contributions of the researchers.

REFERENCES

[1] "Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion." [Online]. Available: https://www.hindawi.com/journals/mpe/2019/9185481/#B3. [Accessed: 30-Dec-2019].

[2] "A Gentle Introduction to Object Recognition With Deep Learning." [Online]. Available: https://machinelearningmastery.com/object-recognition-with-deep-learning/. [Accessed: 03-Jan-2020].

[3] D. Harwath and J. Glass, "Deep Multimodal Semantic Embeddings for Speech and Images," *ArXiv151103690 Cs*, Nov. 2015.

[4] "Multimodal learning - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Multimodal_learning. [Accessed: 03-Jan-2020].

[5] D. F. Harwath, A. Torralba, and J. R. Glass, "Unsupervised Learning of Spoken Language with Visual Context," in *NIPS*, 2016.

[6] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *In CSLDAMT*, 2010, pp. 139–147.

[7] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," *ArXiv150504870 Cs*, Sep. 2016.

[8] D. Harwath and J. R. Glass, "Learning Word-Like Units from Joint Audio-Visual Analysis," *ArXiv170107481 Cs*, May 2017.

[9] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input," p. 17.

[10] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing through ADE20K Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 5122–5130, doi: 10.1109/CVPR.2017.544.

[11] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler, "What Are You Talking About? Text-to-Image Coreference," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3558–3565, doi: 10.1109/CVPR.2014.455.

[12] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping," p. 9.

[13] "Image Caption Generation Using Deep Learning Technique - IEEE Conference Publication." [Online].Available:https://ieeexplore.ieee.org/document/8697360. [Accessed: 05-Jan-2020].

[14] R. Gandhi, "R-CNN, Fast R-CNN, Faster R-CNN, YOLO — Object Detection Algorithms," *Medium*, 09-Jul-2018.[Online].Available: https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e. [Accessed: 05-Jan-2020].

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *ArXiv170306870 Cs*, Jan. 2018.

[16] A. M. Kunnath, "An introduction to implementing the YOLO algorithm for multi object detection in images," *Medium*, 02-Apr-2019. [Online]. Available: https://towardsdatascience.com/an-introduction-to-implementing-the-yolo-algorithm-for-multi-object-detection-in-images-99cf240539. [Accessed: 03-Jan-2020].

[17] Prabhu, "Understanding of Convolutional Neural Network (CNN) — Deep Learning," *Medium*, 21-Nov-2019.[Online].Available: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148. [Accessed: 05-Jan-2020].

[18] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," p. 17.

[19] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, Cham, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[20] W. Havard, L. Besacier, and O. Rosec, "SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO Data Set," *ArXiv170708435 Cs*, doi: 10.18709/PERSCIDO.2017.06.DS80.

[21] Anakha P. J.*, Devika Hari, Rinku Roy, Prof. Joby George 2016. Object Detection and Sentence Generation from Images International Journal of Scientific Research in Science, Engineering and Technology

[22] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit 2013 Isolated speech recognition using mfcc and dtw. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering.

[23] Fang Fang, Hanli Wang, and Pengjie Tang 2018 Image captioning with word level. 25th IEEE International Conference on Image Processing

[24] Simonyan,K.,Zisserman,A.:Verydeepconvolutionalnetworksforlarge-scaleimagerecognition. CoRRabs/1409.1556(2014)

[25] Harwath, David. (2018). Learning spoken language through vision.

[26] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In Proc. NAACL Conference on Human Language Technologies (NAACL-HLT).