Indian Journal of Science and Technology, Vol 9(22), DOI: 10.17485/ijst/2016/v9i22/71557, June 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Obtaining Description for Simple Images using Surface Realization Techniques and Natural Language Processing

C. S. Reddy[1]*, B. Janani[1], S. Arvind Narayanan[1] and E. Mamatha[2]

[1]SASTRA University, Thirumalaisamudram, Tanjore - 613401, Tamil Nadu, India; csreddy@cse.sastra.edu, janani.balasub@gmail.com, sayhitoarvind@gmail.com.
[2]GITAM University, NH 207, Doddaballapur Taluk, Bangalore Rural District, Nagadenehalli, Bangalore – 562163, Karnataka, India; sricsrmax@gmail.com.

Abstract

This paper aims at developing a simple mechanism to deduce corpora pertaining to an image through various computer vision and natural language processing techniques. The output of the vision detection is combined with the sentence formation approach to get the visual content in textual form. Vision detections are smoothed using a number of approaches to prune undesired combination of words that are semantically incorrect. Descriptions are generated based on syntactic trees and Markov Chains and compared for human likeness based on survey. The results of the survey indicate that the descriptions generated with the help of Markov Chains sound more human like. These generated descriptions can be indexed in lucene and image search can be made more efficient bridging the semantic gap.

Keywords: Attributes, Corpora Extraction, Image Detection, Textual Descriptions Generation

1. Introduction

Recounting the pictorial content articulately is a significant facet of communication. Combining vision and language has been tackled by different researches during the recent times. Latest advances in computer vision have improved the vision recognition precision. Nevertheless, there is much room for enhancements in the existing corpora generation schemes. The purpose of the system is to identify the content used to illustrate an image and convert it into descriptive text. This is analogous to the summarizing problem in natural language processing where the most appropriate sentence in a passage that describes the same, is selected as the summary.

A comprehensive application of the obtained corpora would be to improve image search and retrieval with the assistance of the realized textual descriptions by generating text based index for images. The practice of text based retrieval of images is avoided in the prevailing search approaches. But retrieval using text corpora based on the images would significantly improve the search results and user experience. It could also be used to aid visually challenged people. With latest advancements in wearable gadgets, computer vision and language processing can be combined to make a system that would work based on voice query, aiding in navigation as well as identifying the objects in front of visually challenged people. It can also be used to generate auto illustrations. The text generation algorithm can be modified to generate poetry or even stories based on image.

Since it would be impossible to generate textual descriptions for all the available images manually, various learning techniques have been adopted to automatically generate description based on vision outputs for some objects from the PASCAL-VOC Dataset[12]. There are two phases in this generation process. In the first phase, computer vision algorithms are used to identify the various objects, characteristics prominent in the image. In the

*Author for correspondence

next phase, the vision output is smoothed and sentences are generated. The generated sentences are assessed for semantic perfection. Incorrect sentences are pruned.

The main advantage of this approach is that it does not require similar images to produce sentences. Sentences are created from the scratch. However, the main disadvantage is that the computer generated output may not be as eloquent as humans describe. This can be enriched by combining vision outputs with text mined knowledge gained by mining the vast amount of textual information available.

On evaluation, it is observed that the system produces syntactically and semantically correct corpora that sound human-like, for the images. It is also observed that the usage of the results in image retrieval and search tends to give better results.

In the past, captions were formed by obtaining location related information from the image metadata and summarizing related content on the web. The subsequent details give a brief essence on the works relevant to the system in use. This method follows the approach from Kulkarni et al.[1] that creates descriptions for the PASCAL-VOC dataset to accomplish vision recognitions. Their approach builds descriptions with content planning and surface realisation methods to produce corpora that are pertinent to an image.

Previous works in this category makes use of captions from similar images as in[2]. Photo captions are acquired and filtered for noisy results. Image contents are assessed and corresponding descriptions available are selected as descriptions for an image. The procedure in Mitchell et al makes use of word co-occurrence figures obtained by mining the existing descriptions to categorize and neglect incorrect recognitions and descriptions are produced by using syntactic trees. In addition to image description, research has been carried out to generate video descriptions where, in addition to the image object, attribute detection various activity detectors are also used. Human actions are classified hierarchically and these are combined with language generation approaches to get descriptions[3].

The framework proposed by Yao et al. comprises of input images that are disintegrated into visual forms by an image parsing engine. The results are changed into semantic representation. A text generation translates the result into meaningful text. Li et al.[4] have designed a method that produces descriptions by using web-scale n-grams in which there is a phrase selection procedure and which is combined with a phrase fusion approach in order to get textual descriptions. Phrases are generated from vision outputs.

Vision detection results are based on Felzenwalb's Deformable Parts Model[5]. In this algorithm each object is treated as a deformed version of a template. Each image is partitioned into 8x8 pixel blocks. For each block Histogram Orient gradients are computed and features at different resolutions are computed. Detection score is computed as addition of filters and deformation scores. Training data for the models consists of images with labelled bounding boxes taken from PASCAL-VOC challenge. The system also follows Farhadi et al.'s[6] procedure to describe attributes of an image in addition to the recognition method. It incorporates a feature selection method that classifies attributes semantically. It is used to recognise unusual features that could be used as potential adjectives in the sentences.

2. System Overview

An overview of the system in use is given below.

- For an input image, objects in the image are recognized.
- For each object, its attributes are identified.
- The various background elements in the image are identified.
- Prepositions are computed, using a preposition function, from the object detections.
- The detected objects are ordered and grouped.
- Verbs from the collection are hallucinated based on Markov chain rule. Prepositions are also hallucinated in case they cannot be computed from the vision detection.
- Sentences are generated using template based system.
- Triplets and duplets are extracted from the sentences.
- The extractions are checked with the help of a model trained using a large text corpora.

Table 1. Object identification accuracy

Object	Accuracy
Flowers	89%
Building	95%
Shoes	84%

- Only syntactically and semantically correct sentences are taken as descriptions.

3. Image Detections

3.1 Object Detection

Object detectors are basically used to identify what are the objects in the image and where they are located. Detectors trained based on Felzenswalb's Deformable part model and object bank[7] were used to identify various objects in the image. PASCAL-VOC dataset and Image Net dataset were used to train the detectors. Person detectors were trained using the INRIA person dataset. Keywords and bounding boxes for each keyword are extracted from the image. A maximum score is assigned to the detection based on which the most probable nouns are selected. For some objects that are not present in the PASCAL-VOC dataset on which deformable parts algorithm could not be applied, detectors based on Scale Invariant Features are trained. This method summarizes images based on the occurrence of features. It uses weighted product of co-occurrence and then counts the nearest neighbors. A section of our accuracy results have been recorded in Table 1.

$$sim(o_i, p) = \frac{\overrightarrow{o_i}.\overrightarrow{p}}{|\overrightarrow{o_i}| x |\overrightarrow{p}|} = \frac{\sum_{k=1}^{t} w_{k,i} X w_{i,p}}{\sqrt{\sum_{k=1}^{t} w_{k,i}^2} X \sqrt{\sum_{k=1}^{t} w_{i,p}^2}} \quad (I)$$

SIFT vectors for each key point of each object is determined. SIFT vectors consist of points that could be selected under scale changes. For each such point, orientation is identified and descriptors are constructed. They are clustered using K means clustering algorithm. A histogram of features of each object is drawn. K Nearest Neighbours algorithm is used to find the nearest neighbours to find feature matches.

3.2 Attribute Detection

Ability to describe an entity completely depends on the visual attributes associated with it. To identify such visual attributes, classifiers are used. An RBF Kernel SVM trained with images that have bounding boxes and annotations[7] are used. Attributes are not dependent on object category. Therefore it is considered necessary to identify the attributes as they could act as potential text representing images.

3.3. Scene Detection

Linear SVMs trained on low level region features and Histogram of Geometric context were used to classify background scene related information such as grass, sky, etc. The outputs are mapped into probability values and any pixel falling within the detection region is treated as detection. In addition to that, SIFT based classifiers described above are also used to detect some classes of scene related objects, such as buildings.

To match objects and scenes in a simple image, we compute an overall score for each root location value according to the best possible placement of the parts,

$$Match(X_\theta) = \max x_1, x_2, ...x_n score(X_0, X_1, ...X_n) \quad (II)$$

The computing of *match* score also helps in multiple detection of the same object.

4. Corpora Generation

This module creates sentences from 'words' which are the objects and their corresponding attributes that were recognised in the preceding stage. This is essentially the natural language generation problem in NLP. There are numerous techniques to produce sentences. Depending on where the final sentences are going to be used, an ideal approach has to be identified. This paper explores the approach based on syntactic trees and Markov chains as it would be the most suitable method to obtain text for indexing images. Other approaches include the use of N-Gram model and Template based approach, both of which could be used for different applications, the former for auto illustration and the later for generating image excerpts.

4.1 Preposition Function

Propositions for each of the object identified are generated using a simple function. These propositions give the spatial and temporal relationship between the obtained in that step are compared to identify the location of the objects with respect to each other. A set of seven propositions mentioned

in Table 2 are computed with the help of the proposition routine. If the bounding box calculated is wrong, there would be awkward prepositions.

Table 2. List of propositions used

List of propositions
Above, Below, Beside, Near, On, Under, In

4.2 Noun Ordering and Grouping

The data mined descriptions of images are available at[11] and also a large word frequency set is available in[17]. These objects were used to order and group the objects identified. Using the descriptions, among the objects, the one which appears a maximum number of times as root noun is taken as the root noun. When there are a number of objects identified, they have to be ordered in order to be placed in the sentences. Using the data mined objects; the order in which they have to be placed can be computed with the help of the relative occurrence factors. A number of detections of the same object have to also be ignored and grouped. Table 3 depicts how Noun Ordering typically works.

Table 3. Noun ordering data

Before	After
Sofa , person , dog	Person, sofa, dog
Building, Person	Person, building
Car , dog	Dog, car

4.3 Verb and Preposition Hallucination

To identify right verbs either a reliable action detection system has to be developed or computation based on language alone has to be done. This approach follows computation based on language. Verbs are identified using Markov Chain rule. Prepositions are hallucinated based on Markov chain rule in order to avoid awkward prepositions as well as to identify prepositions in case the bounding box calculations could not be obtained or the obtained bounding boxes are wrong. These can be avoided by combining both vision and language based approach in the content planning stage.

4.4 Syntactic Trees

A number of sub trees are generated in this step. First, syntactic sub trees are generated for the <Article, Root> pairs, followed by <Root, Preposition>, and <Preposition, object> trees. <Verb, object> trees are also generated. The root noun is identified from the noun ordering step. For the root noun, the right article to be used is identified with a simple routine. Prepositions are computed using the hand built proposition function previously described. Verbs are either taken from the action recognition output or hallucinated. All the generated sub trees are combined to form fully structured trees. Figure 1. is an example of how Syntactic Tree structure was formed for one of the generated descriptions.

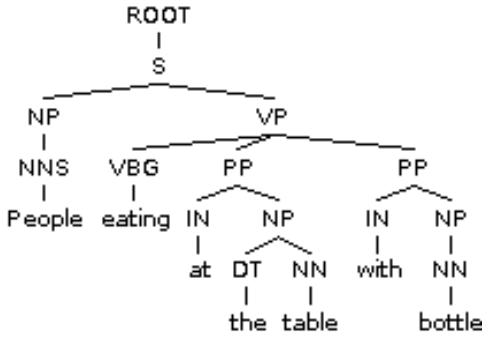

Figure 1. Example final tree.

5. Markov Chains

Sentences are generated based on Markov chain rule and are compared with the syntactic tree sentence generation method. Frequency tables of large corpus are constructed. The root noun in the list of objects identified from the vision system is taken as the initial state. Based on the probability values of the remaining words, Markov based chains are constructed.

$$For\ any\ s, i_0, \ldots i_{n-1} \hat{I}\ S\ and\ any\ n >= 1 \qquad \text{(III)}$$

$$P(A_n = s \mid A_0 = i_0, \ldots A_{n-1} = i_{n-1}) = P(A_n = s \mid A_{n-1} = i_{n-1})$$

This is the basic property of the Markov chains.

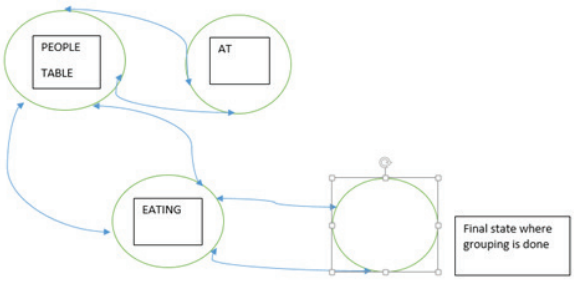

Figure 2. Markov chain rule.

Table 4. Markov Chain states

States	
State 0	People, the table
State 1	eating
State 2	At
State 3	Grouping

Table 3 Talks about the States in Markov chains and their respective generations and Figure 2 shows the corresponding chain formation.

The final state would be the last word in the ordered noun group or a hallucinated or identified verb. These state changes are probabilistic and current state depends on the previous state. For each noun identified, from the probability table, prepositions and verbs are identified. The states are grouped to form a sentence. The sentences generated with the help of Markov Chains are more semantically correct.

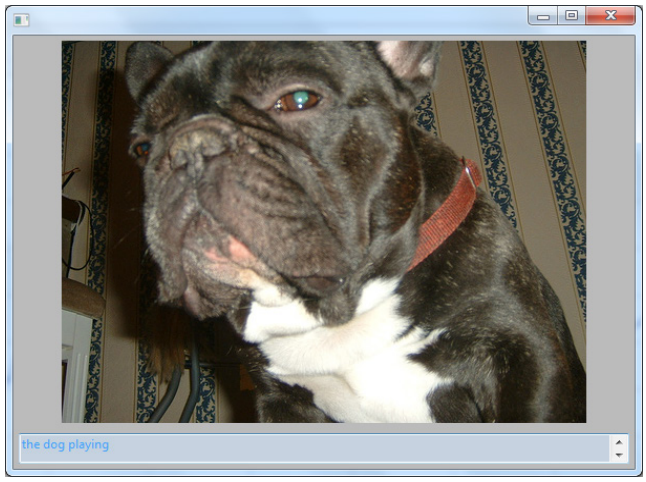

Figure 3. Result

Figure 4. Result

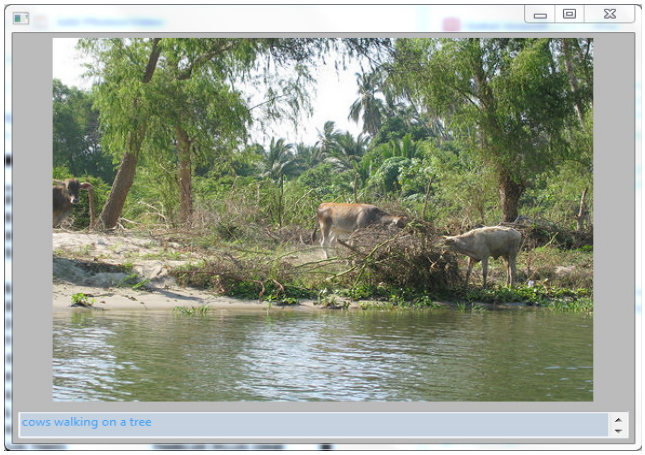

Figure 5. Result

Figure 6. Result

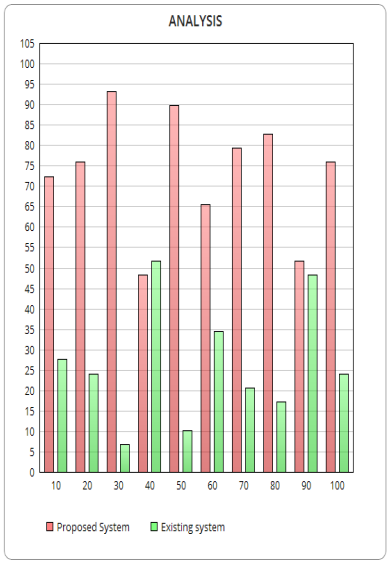

Figure 7. Performance analysis based on survey.

6. Final Grammar Check

Using the Stanford dependency parser, various combinations of words are extracted from the generated sentence, excluding the root noun if the root noun is a person. The various combinations to be extracted are <Verb, Preposition, Object >, <Noun, preposition, Object>, <Object, Preposition, Verb>, <Noun, Verb, Preposition>. These pairs are checked with a model trained based on the data from Google N Grams. From this method, semantically correct sentences can be identified. In case the generated descriptions do not yield semantically correct sentences, the process is repeated with hallucinated verbs and prepositions until a proper sentence that does not contain awkward word positioning is obtained. This surface realisation approach yields semantically correct sentences.

7. Results

The Figures 3 - 7 show some of the descriptions generated by our system for images from the PASCAL VOC dataset.

8. Conclusion

Figure 4 shows a performance analysis of our system built with Markov Chains with the existing system based on a survey. The X-axis denotes a section of our generated results and the Y axis denotes the human likeness score collected from a survey. Thus, the result obtained is syntactically and semantically correct. Although, much of its clarity and accuracy relies up on the computer vision outputs generated. The object detection framework used is the state-of-the art algorithm. But it is not fully accurate. There are some bad results as well. These generated descriptions can be effectively used to index the images for search and retrieval. This bridges the semantic gap to an extent. However, many improvements are needed in the vision algorithm to achieve a significant improvement in the semantic gap problem.

9. Future Work

In future, the currently used computer vision algorithms are planned to be replaced by neural networks for more accuracy. In combination with object detection, scene, and attributes present in the Image Net dataset, more attributes, face recognition and better activity detectors will be used to generate better descriptions which can be effectively utilised in solving the semantic gap problem to an extent.

10. Acknowledgements

We would like to thank the Dean of School of Computing for supporting us with this effort.

11. References

1. Kulkarni G, et al. Baby Talk: Understanding and Generating Simple Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013 Dec; 35(12):2891-2903.
2. Ordonez V, Kulkarni G and Berg TL. Im2text: Describing images using 1 million captioned photographs. Proc. NIPS 2011. 2011.
3. Krishnamoorthy N, Malkarnenkar G, Mooney R. Generating natural-language video descriptions using text-mined knowledge-Procedings of AAAI, 2013.
4. Li S, Kulkarni G, Berg TL, Berg AC and Choi Y. Composing simple image descriptions using web-scale n-grams. Stroudsburg, PA, USA: Association for Computational Linguistics: In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL '11). p. 220-228.
5. Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Computer Vision and Pattern Recognition. 2008. CVPR 2008. IEEE Conference on Computer Vision and Pattern Recognition. 2008 June; 1(8):23-28.
6. Farhadi A. Endres I, Hoiem D, Forsyth D. Describing objects by their attributes. IEEE Conference on Computer Vision and Pattern Recognition. 2009 20th to 25th June; p. 1778-85.
7. Li LJ, Su H, Xing AP, Fei-Fei L. Object bank: A high-level image representation for scene classification and semantic feature sparsification. Advances in Neural Information Processing Systems.
8. Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daume H III. Midge: generating image descriptions from computer vision detections. Avignon, France: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012 April 23-27.
9. Katare A, Mitra SK, Banerjee, Asim. Content Based Image Retrieval System for Multi Object Images Using Combined Features. ICCTA '07 International Conference

on Computing: Theory and Applications 2007. 2007 March 5-7; 595(599).

10. Ordonez V, Kulkarni G, Berg TL. Im2Text: Describing Images Using 1 Million Captioned Photographs. Neural Information Processing Systems(NIPS). 2011.
11. Benjamin Z Yao, Yang X, Lin L, Lee MW and Zhu S. 2010. Proceedings of IEEE, I2T: Image parsing to text description. 2010; 98(8):1485-1508.
12. Mitchell M, Dunlop A, Roark B. Semi-supervised modeling for prenominal modifier ordering. Portland, Oregon: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers. 2011 June 19-24.
13. Everingham M, Gool LV, Christopher KI Williams, Winn J, Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision. 88(2):303-38.
14. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - 2005 June 20-26; 1:886-893.
15. Flickr. 2011. Date accessed 1.Sep.11: Available from: http://www.flickr.com.
16. Michel JB, et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science, Published online ahead of print. 2010.
17. Marneffe MC, MacCartney B and Christopher D Manning. In LREC 2006: Generating Typed Dependency Parses from Phrase Structure Parses. 2006.
18. Javubar K Sathick, Jaya A. Natural Language to SQL Generation for Semantic Knowledge Extraction in Social Web Sources. Indian Journal of Science and Technology. 2015 Jan; 8(1):1-10.