Review on Textual Description of Image Contents

Vasundhara Kadam#1, Ramesh M. Kagalkar*2

#ME Second Year Student

Dr D Y Patil School of Engineering and technology, Lohegaon, Pune, India

*Research Scholar and Assistant Professor

Dr D Y Patil School of Engineering and technology, Lohegaon, Pune

Abstract—*Visual image relation with visually descriptive language is a major challenge for computer vision specifically becoming additional relevant as recognition as well as detection techniques are beginning to work. This paper reviews on techniques that are used for image description such as associations between objects present in that image. Additionally, paper presents an approach to automatically make natural language descriptions from images shortly. This proposed system consists of two parts called content planning and surface realization. The first part, content planning, smooths the output of computer vision-based recognition and detection algorithms with statistics extracted from large groups of visually descriptive text to define the best content words to use to define an image. The another step, surface realization, selects words to build natural language sentences based on the projected content and overall statistics from natural language.*

Keywords — *Computer vision, image description generation, content planning, surface realization.*

I. Introduction

Automatic description of images in natural language is an intriguing, but complex Artificial intelligence (AI) task, needing accurate computational visual recognition, wide-ranging world knowledge, and generation of natural language. Natural language, i.e. whether written, spoken, or typed, makes up much of human communication. A significant extent of this language defines the visual world either directly around us otherwise in images as well as video. Visual image association with visually descriptive language is a challenge for computer vision that is becoming much more relevant as recognition as well as detection methods are beginning to work. Studying natural language and particularly how people describe the atmosphere around them can support us to better understand the visual world. It can help us in order to generate natural language in the quest that describes this world in a human manner.

There is a huge quantity of visually descriptive text presented both closely related with images in descriptions and in pure text documents. Studying such type of language has the potential to deliver (i) training data in order to understand how people describe the world or environment, as well as (ii) more common knowledge around the visual world indirectly encoded within human language.

A better understanding of what data it is necessary to extract from an image so as to decide on applicable descriptive language may cause new or additional observer-focused goals for recognition. it is subtle, however many factors distinguish the challenge of taking images as input then generating descriptions of natural language from several alternative tasks in computer vision. As examples, once forming descriptive language, people go beyond simply listing which objects are present in an image this is correct even for images having very low-resolution and for very brief exposure to images. In each of those settings and in language normally, individuals include specific data describing not only scenes, however specific objects, their relative locations, and modifiers adding additional data regarding objects. Mining the absolutely huge amounts of visually descriptive text accessible in different library collections and on the web normally makes it possible to get what modifiers people use to describe objects and what prepositional phrases are used to describe relationships among objects. These are often used to select and train computer vision algorithms to recognize these constructs in images. The output of the computer vision process may be "smoothed" using language statistics and so combined with language models during a natural language generation method.

Natural language generation constitutes one among the basic research issues in natural language process (NLP) and is core to a large vary of natural language processing applications like Machine Translation (MT), text summarization, dialogue systems, and machine-assisted revision. Despite considerable progression within the last years, natural language generation still remains an open analysis drawback. Most previous work in natural language processing on automatically generating captions or descriptions for images relies on retrieval and summarization. For example, Aker and Gaizauskas [8] believe GPS metadata to access relevant text documents and Feng and Lapata [11] assume relevant documents are provided. The method of generation then becomes one among combining or summarizing appropriate documents, in some cases driven by keywords calculable from the image content [11]. From the computer vision perspective these techniques could be analogous to first recognizing the scene shown in an image then retrieving a sentence based on the scene type. From the computer vision community, work has considered matching an entire input image to a database of images with captions [10], [6]. The caption of the most effective matching image will then be used for the input image. These approaches mentioned that sample directly from human written text might produce additional natural sounding, albeit probably less directly relevant or descriptive output. Additionally to reviewing the generation approach of Kulkarni et al. [4] and presenting a replacement surface realization strategy using additional versatile optimization.

This paper explores techniques to profit from each of those possible sources of information. The primary type of textual information is exploiting as a previous to modulate global inference above computer vision-based objects recognition, appearance characteristics, and background regions. The second form of language data is exploited to convert the resulting keyword-based predictions into complete and human-like natural language descriptions. Additionally to the direct outputs of system automatically generated natural language descriptions for images there also are variety of possible connected applications. These include improving accessibility of images for the visually impaired and making text-based indexes of visual knowledge for improving image retrieval algorithms. Additionally, work is in line with an additional general research direction toward learning visually descriptive text and delving deeper into the association between images and language that has the potential to suggest new directions for analysis in computer vision. The projected approach is comprised of two stages. Within the first, content planning, the typically noisy output of algorithms presented for computer vision recognition is smoothed with statistics collected from visually descriptive natural language. Once the content to be employed in generation is chosen, subsequent stage is surface realization, finding words to explain the chosen content. Once again text statistics are used to select surface realization that is more almost like constructions in normally used language.

The rest of paper is divided into some sections as follows: Section II gives the essential background. Section III addresses feedback session overview. Section IV introduces the mapping concept among feedback session and pseudo documents. Section V describes previous techniques used for clustering and finally section VI concludes the summary of paper.

II. Literature Review

This section briefly reviews some of the most relevant work.

A. *Integrating Words and Pictures*

J. Sivic et al. [25] investigate the problem of automatically labelling faces of characters in TV or show material with their names, by means of only weak supervising from automatically aligned subtitle and script text. Authors designed a method extending the coverage importantly by the detection in addition to recognition of characters in profile views additionally with (i) seamless following, integration and recognition of profile and frontal detections, and (ii) a character specific multiple kernel classifiers which able to learn the features best able to discriminate between the characters.

Li-Jia Li et al. [23] presents an automatic dataset col-lecting and model learning approach that uses object recognition techniques in an incremental methodology. It mimics the human learning method of iteratively accumulating model information and image examples. Authors adapt a non-parametric graphical model and propose a progressive learning framework.

B. *Learning Models of Categories or Relationships*

Chaitanya Desai et al. [17] introduce a unified model for multi-class object recognition that casts the issue as a struc-tured prediction task. Instead of predicting a binary label for every image window independently, their model at the same time predicts a structured labeling of the whole image. This model learns statistics that capture the spatial arrangements of various object categories in real images, each in terms of those arrangements to suppress through non-maxima suppression (NMS) and those arrangements to favor through spatial co-occurrence statistics.

An associated body of work on image parsing and object detection, learns the spatial relationships between labeled components either detections or regions. These relationships were used as contextual models to enhance labeling accuracy; however the spatial relationships themselves were not considered outputs in their own right. An approach for learning a discriminative model of object categories, incorporating texture, layout, and context data is presented in [24].

A. Torralba et al. [14] presented a probabilistic framework for encoding the relationships between context and object properties.

C. *Object Attributes*

Ali Farhadi et al. [18] introduced a novel feature selection methodology for learning attributes that generalize well across categories. Mainly, they presented an attribute-centric approach for learning object attributes.

Neeraj Kumar et al. [21] present two novel strategies for face verification. The first methodology “attribute” classifiers uses binary classifiers trained to recognize the presence or absence of describable aspects of visual appearance (e.g., gender, race, and age). The second methodology "simile" classifiers removes the manual labelling needed for attribute classification and in its place learns the similarity of faces, or regions of faces, to specific reference people.

Christoph H. Lampert et al. [22] tackle the problem of object classification once training and test categories are disjoint, by introducing classification of attribute-based. It performs object detection based on a human-specified high-level description of the target objects instead of training images.

Tamara L. Berg et al. [9] explores automatic discovery of attribute vocabularies and visual representations learning from unlabelled image and text information on the web. This methodology is able to dependably find and rank potential attribute phrases according to their visualness a score associated with however strongly a string is correlated with some aspect of an object's visual appearance.

Josiah Wang et al. [26] investigate the task of learning models for recognition of visual object from natural language descriptions alone. The approach contributes to the recognition of fine-grain object categories, like animal and plant species, where it may be tough to collect several images for training, however where textual descriptions of visual attributes are readily available.

D. Describing Images

Vicente Ordonez et al. [6] develop and demonstrate automatic image description strategies using a giant captioned photograph collection. They additionally develop strategies incorporating many states of the art, however fairly noisy, estimates of image content to produce even additional pleasing results.

Ali Farhadi et al. [10] presented methodology that computes a score linking an image to a sentence. This score is used to attach a descriptive sentence to a given image, or to obtain images that illustrate a given sentence. The score is attained by equating an estimate of which means obtained from the image to one obtained from the sentence.

P. Kuznetsova et al. [3] present a holistic data-driven approach to image description generation, exploiting the huge quantity of (noisy) parallel image information and associated natural language descriptions available on the online.

Li et al. [5] focus on introducing creativity in sentence construction. They presented a straightforward yet effective method to automatically comprise image explanations given computer vision based inputs and using web-scale n-grams.

Finally, Yang et al. [14] also compose descriptions in a bottom up fashion, detection objects and scenes, then using text information to "hallucinate" verbs for objects. Descriptions are then composed in an HMM framework.

E. Describing Videos

Abhinav Gupta et al. [19] presented an approach to learn a visually grounded storyline model of videos directly from weakly labelled information. The storyline model is represented as an AND-OR graph, a structure which will compactly encode storyline variation across videos. The edges within the AND-OR graph correspond to causal relationships that are represented in terms of spatio-temporal constraints.

Sonal Gupta and R. J. Mooney [20], [12] explores how closed captions that naturally accompany many videos will act as weak supervision that allows automatically collecting 'labelled' information for activity recognition. They additionally present a novel caption classifier that uses further linguistic information to work out whether or not a selected comment refers to an ongoing activity.

III. Proposed Framework

An overview of proposed scheme can presented as shown in figure 1 as well as explained as follows:

1. Detectors are used to detect things (e.g., bus, car, bird, person, etc.) and stuff (e.g., grass, trees, road, water, etc.). The proposed framework will designate these as things and stuff, or as a group of objects.
2. Every candidate object (i.e. either thing or stuff) region is processed by a set of attribute classifiers.
3. Every pair of candidate regions is processed by prepositional relationship functions.
4. A conditional random field (CRF) is constructed that incorporates the unary image potentials calculated by step 1-3, with higher order text-based potentials calculated from large text corpora.
5. A labelling of the graph is predicted.
6. Sentences are generated.

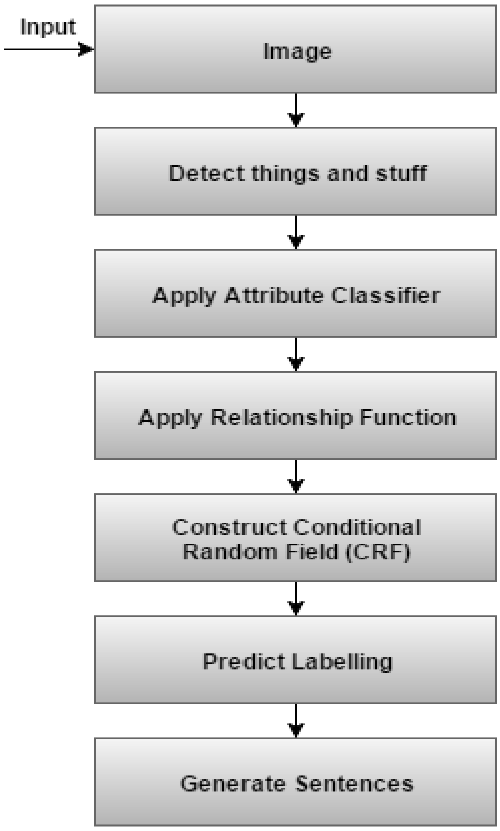
Input
Image
Detect things and stuff
Apply Attribute Classifier
Apply Relationship Function
Construct Conditional Random Field (CRF)
Predict Labelling
Generate Sentences

Fig. 1 System Architecture

IV. Proposed System Description

One major potential practical advantage of the approach presented in this paper is that it will generate descriptions without requiring related text or similar images with descriptions. Instead, it builds a caption for an image in a bottom up fashion, starting from what computer vision systems recognize in an image and then constructing a novel caption around those predictions, using text statistics to smooth these (sometimes) noisy vision predictions. However, the downside of such an approach is that descriptions are created entirely from scratch. The alternative approaches mentioned above [10], [6] that sample directly from human written text could produce additional natural sounding, albeit probably less directly relevant or descriptive output. These and different competing desirable traits (e.g., accuracy to content, and naturalness of expression) in natural language description create challenges for analysis. Additionally to reviewing the generation approach of Kulkarni et al. [4] and presenting a new surface realization strategy using additional flexible optimization, this paper presents extensive novel evaluations of the generated sentences of this system and evaluations comparing the generated sentences with those from competitory approaches. Evaluations are performed either automatically by measure similarity of generated sentences to reference examples written by humans, or by directly asking humans that of two sentences could be a higher description for an image.

A. Content Planning

A conditional random field (CRF) is used to predict a labeling for an input image. Nodes of the CRF correspond to several types of image content: (i) objects things or stuff, (ii) attributes that modify the appearance of an object, and (iii) prepositions that refer to spatial relationships between object-object pairs (including things and stuff). To predict the most effective labeling for an input image graph (both at test time and during parameter training) planned system utilizes the sequential tree reweighted message passing (TRW-S) algorithm.

B. Surface Realization

The output of CRF could be a predicted labelling of the image. This forms the content need to encode in surface realization step, generation of the final natural language descriptions. This labelling encodes three types of information: objects present within the image (nouns), visual attributes of these objects (modifiers), and spatial relationships between objects (prepositions). The projected system presents three generation techniques for producing a surface realization. The first is based on decoding using n-gram language models second id additional flexible ILP-based optimizations which will handle a wider vary of constraints on generation, third is template-based approach [1].

V. Conclusions

This paper presented a survey of strategies on textual description of Image Contents as well as related work, including: using word and image information jointly for labelling images, learning models of categories, attributes, or spatial associations from data, as well as methods to compose descriptions for images. Moreover paper presents a system to automatically produce natural language descriptions from images.

Acknowledgment

The authors would like to thank the researchers and publishers for creating their resources obtainable and professors for their guidance. We are thankful to the Mr. Ramesh M kagalkar for his valuable guidance. Finally, we would like to extend a heartfelt gratitude to friends and family members.

References

[1] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Alexander 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, December 2013.

[2] P.F. Felzenszwalb, R.B. Girshick, and D. McAllester, “Discriminatively Trained Deformable Part Models, Release 4,” http://people.cs.uchicago.edu/pff/latent-release4/, 2012.

[3] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, and Y. Choi, “Collective Generation of Natural Image

Descriptions," Proc. Conf. Assoc. for Computational Linguistics, 2012.

[4] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg, "Babytalk: Understanding and Generating Simple Image Descriptions," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2011.

[5] S. Li, G. Kulkarni, T.L. Berg, A.C. Berg, and Y. Choi, "Composing Simple Image Descriptions Using Web-Scale n-Grams," Proc. 15th Conf. Computational Natural Language Learning, pp. 220-228, June 2011.

[6] V. Ordonez, G. Kulkarni, and T.L. Berg, "Im2text: Describing Images Using 1 Million Captioned Photographs," Proc. Neural Information Processing Systems), 2011.

[7] Y. Yang, C.L. Teo, H. Daume, and Y. Aloimonos, "Corpus-Guided Sentence Generation of Natural Images," Proc. Conf. Empirical Methods in Natural Language Processing, 2011.

[8] A. Aker and R. Gaizauskas, "Generating Image Descriptions Using Dependency Relational Patterns," Proc. 28th Ann. Meeting Assoc. for Computational Linguistics, pp. 1250-1258, 2010.

[9] T.L. Berg, A.C. Berg, and J. Shih, "Automatic Attribute Discovery and Characterization from Noisy Web Data," Proc. European Conf. Computer Vision, 2010.

[10] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D.A. Forsyth, "Every Picture Tells a Story: Generating Sentences for Images," Proc. European Conf. Computer Vision, 2010.

[11] Y. Feng and M. Lapata, "How Many Words Is a Picture Worth? Automatic Caption Generation for News Images," Proc. Assoc. for Computational Linguistics, pp. 1239-1249, 2010.

[12] S. Gupta and R.J. Mooney, "Using Closed Captions as Supervision for Video Activity Recognition," Proc. 24th AAAI Conf. Artificial Intelligenc, pp. 1083-1088, July 2010.

[13] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting Image Annotations Using Amazon's Mechanical Turk," Proc. NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk, 2010.

[14] A. Torralba, K.P. Murphy, and W.T. Freeman, "Using the Forest to See the Trees: Exploiting Context for Visual Object Detection and Localization," Comm. ACM, vol. 53, pp. 107-114, Mar. 2010.

[15] M.-C. de Marnee and C.D. Manning, Stanford Typed Dependencies Manual, 2009.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[17] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative Models for Multi-Class Object Layout," Proc. 12th IEEE Int'l Conf. Computer Vision, 2009.

[18] A. Farhadi, I. Endres, D. Hoiem, and D.A. Forsyth, "Describing Objects by Their Attributes," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[19] A. Gupta, P. Srinivasan, J. Shi, and L.S. Davis, "Understanding Videos Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Videos," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[20] S. Gupta and R. Mooney, "Using Closed Captions to Train Activity Recognizers that Improve Video Retrieval," Proc. IEEE Computer Vision and Pattern Recognition Workshop Visual and Contextual Learning from Annotated Images and Videos, June 2009.

[21] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar, "Attribute and Simile Classifiers for Face Verification," Proc. 12th IEEE Int'l Conf. Computer Vision, 2009.

[22] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[23] L.-J. Li and L. Fei-Fei, "OPTIMOL: Automatic Online Picture Collection via Incremental Model Learning," Int'l J. Computer Vision, vol. 88, pp. 147-168, 2009.

[24] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context," Int'l J. Computer Vision, vol. 81, pp. 2-23, Jan. 2009.

[25] J. Sivic, M. Everingham, and A. Zisserman, ""Who Are You?" Learning Person Specific Classifiers from Video," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[26] J. Wang, K. Markert, and M. Everingham, "Learning Models for Object Recognition from Natural Language Descriptions," Proc. British Machine Vision Conf., 2009.