Vol.4,No.1, July, 2015
190 - 196,ISSN:0975-5446] [VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY]

Integrating Computer Vision and Natural Language Processing : Issues and Challenges

SHUKLA Deepika
Computer sc. and Eng. Dept
Institute of Technology, Nirma University,Ahmedabad(India)
deepika.shukla@nirmauni.ac.in

DESAI Apurva A.
Department of Computer Sc. and Information Technology
Veer Narmad South Gujarat University,Surat(India)
aadesai@vnsgu.ac.in

Abstract

Computer vision and natural language processing (NLP), although both being an active machine learning research areas in itself, is currently treated as two separate area of research and work going on in one area is found to be less benefitted from the other one. But recently many researchers have started exploring the possibility that whether the text accompanying images or videos can somehow be utilized to make the computer vision simpler and efficient. At the other end, efforts are made to find out that, how computer vision can help the tasks inherent to NLP, like language learning and to make out the meaning of words and sentences using visual information. Moreover, there are also a varied and important real-world applications that require integrating vision and language.
In this study feasibility of integrating computer vision and natural language processing has been explored from the existing literature on the topic. Also an attempt is made to suggest the futuristic applications where computer vision can be used to aid natural language processing and also the way in which natural language processing can be helpful in computer vision.

Keywords : Computer vision , Natural language processing , Machine learning, Image analysis, Image understanding.

1. Introduction:

Human beings use their vision capabilities and language to perceive the world around them and to communicate it to others. Given an image they can easily give a language description of it and similarly when something is described in language, they can very well create an image out of it. For human being this is a mundane activity, but imbibing these capabilities in machine, is a challenging task. It requires the knowledge of research areas like computer vision and natural language processing. Although, both the fields have emerged from Artificial Intelligence and are today active area of research in themselves, is treated by the

researchers of each field in isolation. Research taking place in one of the field seems to be less benefitted from the other one. Both the fields have flourished separately and have given exciting applications but majority of current applications have not exploited the visual and textual information jointly. However, there exists many real world problems which require knowledge and expertise from both the field as the co-occurrence of visual and textual data is natural and significant and easily available; for example subtitles in the videos, tagged images on social networking sites etc. Also with the tremendous growth of visual and textual data on web as well as private repositories has generated requirement of searching organizing and mining this data for varied reasons. So it is natural and mandatory to explore the possibilities, how the integration of both the field can be achieved and where it would be beneficial for the resultant application. For the completeness of the content, we explain further, both the tasks i.e computer vision and natural language processing and activities involved in it in the following sub-sections. Further, the paper reviews previous research on the efforts and systems created using the concept of integrating computer vision and natural language processing. We describe the research issues and challenges those would be encountered while bridging the gap between computer vision and NLP. The remainder of the paper is organized as follows. Section 2 discusses related work happening for the area. Section 3 is devoted to throw some light on the research issues and challenges inherent in this kind of integration. Section 4 describes the applications where the concept has been exploited. Section 5 discusses some of the application which as per our view the concept can be deployed and systems can be made. Finally Section 6 is presents the conclusion of this study.

1.1 Computer Vision

At the broadest level computer vision can be seen as enabling machines to see and perceive the world the way and with the ease, in which we humans do. In order to do so the field includes activities like processing and understanding images or visual data. The need of incorporating vision capabilities in the contemporary applications being developed and deployed in industry have increased at a noticeable pace. Applications like object detection and recognition, surveillance systems, activity recognition, image rendering are in great demand. To fulfil this need, tremendous research is going on in research organizations and academia.

There could be many activities involved in a computer vision task and moreover these activities generally are application dependent. However, Figure-1 explains computer vision in principle and typical activities involved in any generic computer vision task.

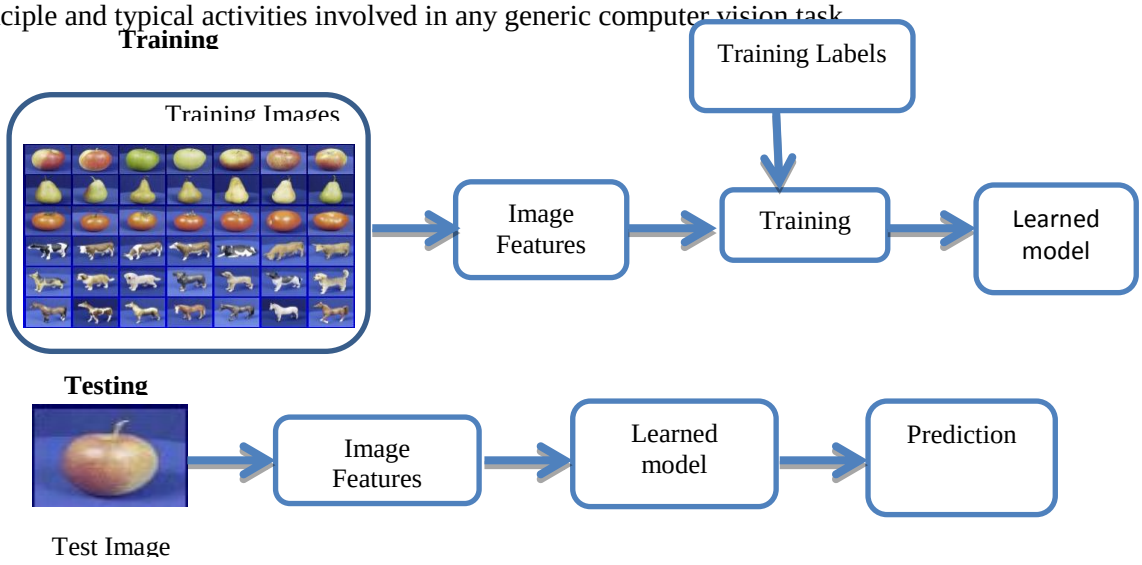

Figure-1 Activities involved in a typical computer vision system

In addition to these the images are to be pre-processed for noise reduction, orientation alignment or some kind of enhancement is done to make the image more suitable for vision task.

1.2 Natural Language processing

The other area that is considered is NLP. At the very core, NLP means enabling computers to derive meaning from spoken or written text in natural language input. The spectrum of NLP also includes generating sentences in natural languages by computers just like we humans do. Figure 2 shows the list of tasks that are performed in a typical Natural Language Application.

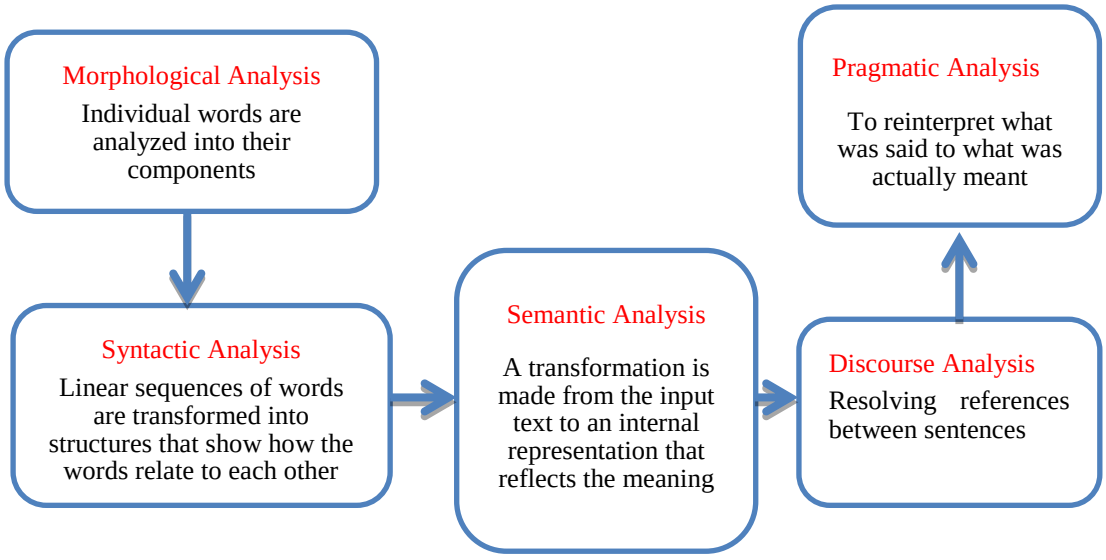

Figure 2: Stages of Natural Language Processing(NLP)

1.2 Scope of Integration of Computer Vision and NLP

i.) A huge number of tagged images and video exist on web, especially social networking sites, which can be used to develop databases of images which can be in turn used to train object recognition and activity recognition systems.
ii.) A picture is equivalent to 1000 words. Images can be generated from textual data which is readily available on web which can be mined to convert them into images which can further be supplied to vision systems.

2. Literature Review

As mentioned in previous sections that computer vision and natural language processing, despite being originated from common area of Artificial Intelligence, regardless of few instances and eventual relationship, are treated as distant fields of research in research fraternity. Here, in this section we discuss and try to justify the places where two of them can be integrated and one can be used to aid the tasks carried out in other and where such instances are found in literature. This integration can be done at two levels; especially when NLP aiding vision is concerned. In [2], it has been shown that the integration can be achieved with two aspects (i) Many types of linguistic knowledge or knowledge extracted from text can be used to aid computer vision. (ii) Transferring algorithmic techniques from one domain to another i.e either algorithms which are initially developed for language processing is successfully is/can be adopted to vision tasks or algorithms developed for doing vision tasks are now being used to do language processing.

2.1 Linguistic knowledge or knowledge extracted from text aiding computer vision

Many approaches exist in literature, where statistics gained from accompanying text is used to get complementary data related to visual recognition task. Statistics can be gathered for occurrence of Subject-Verb-Object combinations and their joint probabilities. Mooney and Motwani in [3] have used such statistics for activity recognition where probability estimates of subject-verb-object (SVO) combinations is exploited to aid the task of activity recognition instead of using purely visual information from the image. In activity recognition task, usually set of activity classes to be recognized are explicitly provided whereas in their work usage of probabilities of SVO combinations is done to discover automatically the set of activities from textual descriptions. Similar statistics have been generated in [1] for giving sentential description of images and predict the scenes from objects and verbs. The authors have used a language model trained from English Gigaword corpus to obtain their estimates; together with probabilities of co-located nouns, scenes and prepositions. They have demonstrated that, predicting the most likely nouns, verbs, scenes and prepositions that make up the core sentence structure from still images directly is very unreliable. Few authors have adopted similar approach for videos also [4]. The authors in [4] have used SVO triplet automatically mined from web-scale text corpora and have gained enhancement in activity identification by providing contextual information to SVO triplet selection algorithm. One can use object-object co-occurrence data extracted from text to acquire knowledge that could aid joint recognition of multiple objects in an image. For example; it is more likely that an 'Apple' exist in the same image which contains 'Banana' rather than in an image which contains an 'Aeroplane'. Similar relationship can be extracted and learned for the series of events also [11, 12]. In this work, events are learned as the logical consequence of each other and thus learn the temporal relationship between pair of events and prove that the existence of one event can support the existence of another event which can further facilitate the task of activity recognition.

2.2 Instances of algorithmic adoption and integration:

BoW Model: Bag of Words model is representing a text document or a sentence written in natural language, as set of words, not taking into consideration its grammar or the order in which these words occur in the original text. The frequency of occurrence of each word is calculated and then used for various language processing tasks. Initially, the technique was developed in the field of natural language processing (NLP) for text document analysis, but in the near past, it has been successfully adapted by the vision community. The analogous term BoF(Bag of Features), is used to represent the approach. Similar to BoW model, here image is represented as order less collections of local features of Image. Recently many authors have used the approach to accomplish various vision related tasks like image classification, object recognition (e.g [15]), texture recognition [16] , image annotation , image retrieval etc.

Usage of deep architectures for deep learning of features: Deep learning neural networks have been adopted well to solve varied vision problems which in turn have brought tremendous improvement in the performance of image analysis results over the last few years. Recent developments in machine learning known as 'Deep Learning' have made it possible to learn features in an unsupervised manner directly from data instead of handcrafting them explicitly. The approach has helped vision tasks particularly object recognition and large scale image classification [17] greatly, thereby enabling effective capturing of low-level as well as middle level cues of object to be recognized. In the recent past the technique is widely applied to the area of NLP. Socher et al in [18] demonstrates the same. The methods developed by them are based on deep learning and they have extended general deep learning ideas by introducing recursion and computing representations for grammatical language structures.

Databases: Imagenet Vs Wordnet : ImageNet is an image database inspired in principle by the organization of the 'WordNet' hierarchy, in which each node of the hierarchy is depicted

by hundreds and thousands of images. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet, majority of them are nouns (80,000+). Similarly, in 'ImageNet', around 1000 images are used to illustrate each synset and currently total number of non-empty synsets available on Imagenet are 21841 [20]. Also, synsets are created for almost 200 categories of objects.

3. Challenges and issues in Computer Vision and NLP

One of the biggest challenges the computer vision field is facing is related to creating the training and testing dataset which can be used to train the vision system. Traditional machine learning for both computer vision requires manually annotating images, video, text, or speech with detailed labels, parse-trees, segmentations, etc. This task of annotating has to be done by human experts and turns out to be mammoth task considering huge number of objects existing in the real world which are to be recognized for various application. Moreover, these objects are visible as entirely different object if the image taking condition like illumination, viewpoint etc. are changed or the position, scale or orientation of the object in the image changes. i.e; with any change in the object or image taking condition it results into entirely new image. Now, annotating such a large number of classes of objects and each with huge number of variations turns out to be a herculean task. As mentioned earlier, the task of labelling or manual segmentation is performed by human experts and it is not so easy to get the human experts accomplishing this effectively, correctly and efficiently. At the other extreme, NLP is successful remarkably as far as morphological analysis and syntax is concerned [6] but if one considers other phases of language understanding, still many things are to be achieved. Applications are still in great demand which handles the issue of semantic analysis, word sense disambiguation or identifying meaning of tenses to temporal objects. Challenges are there in the pragmatic analysis also. A simple declarative sentence stating a fact, "it is sunny" for example is not only a statement of fact but also serves some communication function. The function may be to inform, to mislead about fact or speaker's belief about fact, to draw attention, to remind previously mentioned event or object related to fact, etc. So, the pragmatic interpretation seems to be open ended. Also these challenges become more crucial considering the fact that there exist huge number languages across nations. Most of the existing work is done for 'English' language only. Looking the above scenarios it can be said that methods that integrate language and vision hold the promise of greatly bridging this gap by using naturally co-occurring text and images/video and can mutually complement and supervise each other.

4. Applications

This section lists the applications where either the visual data is used for language understanding or natural language processing of textual data is performed for accompanying vision task.

4.1 Application areas where the concept is already in use or is being researched.

a. Robotics : In the field of robotics, the concept is used widely. The robots needs to perceive their surrounding from more than one way of interaction. Moreover, spoken language and natural gestures are more convenient way of interacting with a robot for a human being, if at all robot is trained to understand this mode of interaction. From the human point of view this is more natural way of interaction as compared to data gloves and all. In work mentioned in [8,10,14] can be considered as examples of such approach where authors have used speech and gesture recognition for interacting with the mobile robots.

b. Recognizing objects from images from their textual description especially where it is difficult to acquire many training images. For example studies related to wild life of rare species of animals and plants. In such cases, usually, textual description is readily available

on web or other repositories. One such effort is mentioned in [13], to recognize particular specie of butterfly learning models from online nature guide.

c. Project VITRA (Visual Translator): is concerned with the development of knowledge-based systems for natural language access to visual information [9]. In the project different domains of discourse and communicative situations are examined with respect to natural access to visual information. Various situations are examined like:

(i) Answering questions about observations in traffic scenes.
(ii) Generating running reports for short sections of soccer games
(iii) Describing routes based on a 3-dimensional model of the University Campus.

5. Suggested Applications

The concept of integrating computer vision and NLP has been used in many applications. Still in this sub-section we suggest certain more applications where the concept can be deployed.

a. Designing: The integration of visual data and text can be incorporated into designing of homes/home decor as well as clothes, jewellery or any such item. The customer can explain the requirement verbally or in written form and this description can be converted to image which can be shown to the customer for better visualization.
b. Generating description of medical images which can be used by doctors.
c. Automatic caption generation for news images or generating sub-titles for movies
d. Converting sign language to speech or text.
e. Making a system which sees the surrounding and gives a spoken description of the same can be used by blind people.
f. Making systems which can convert spoken content in form of some image which may assist to an extent to people which do not possess ability of speaking and hearing.

6. Conclusion

From the literature review it can be concluded that it is very intuitive to use natural language processing to ease the task inherent to computer vision rather than less instances we were able to get where visual information was used to perform the tasks inherent to NLP like understanding the meaning of the sentences written in the images of text document. The work that can be considered as the first step in this direction is OCR (optical character recognition) but it limits only to the recognition of content from the text document images. Once the OCR tasks are done then whole semantic understanding and pragmatic analysis needs to be addressed and can be considered as a novel direction of research. Additionally, the systems performing language processing for other languages would be great contribution. Another direction of work that is researched is; given images or video, from it the natural language description is generated. This kind of work can be very useful in case of sub-title generation or caption generation and systems can be developed which could assist differently able people like blind or deaf and dumb. Finally, from the study we conclude that integrating computer vision and natural language processing has long way to go and holds very promising future.

Acknowledgement

We acknowledge UGC for Special Assistance Program (SAP) for Natural Language Processing and Data Mining (file number is F.3-48/2011) under which this work has been done.

References

[1] Yang, Yezhou, et al.Corpus-guided sentence generation of natural images. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[2] cs.washington.edu/mssi/2013/mooney-msr-uw-2013.pdf
[3] Tanvi S. Motwani, Raymond J. Mooney ,Improving Video Activity Recognition using Object Recognition and Text Mining,In Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012), pp. 600--605, August 2012.
[4] Niveda Krishnamoorthy and Girish Malkarnenkar and Raymond J. Mooney and Kate Saenko and Sergio Guadarrama, Generating Natural-Language Video Descriptions Using Text-Mined Knowledge, In Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI-2013), pp. 541--547, July 2013
[5] http://www.cost.eu/COST_Actions/ict/Actions/IC1307
[6] http://language.worldofcomputing.net/nlp-overview/open-problems-in-natural-langauge-processing.html
[7] Thomason, Jesse, et al. "Integrating language and vision to generate natural language descriptions of videos in the wild." Proceedings of the 25th International Conference on Computational Linguistics (COLING), August. 2014.
[8] Perzanowski, Dennis, Alan C. Schultz, and William Adams.,Integrating natural language and gesture in a robotics domain. Intelligent Control (ISIC), 1998. Held jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA), Intelligent Systems and Semiotics (ISAS), Proceedings. IEEE, 1998.
[9] Herzog, Gerd, and Peter Wazinski. "Visual translator: Linking perceptions and natural language descriptions." Artificial Intelligence Review 8.2-3 (1994): 175-187.
[10] Yang, Yezhou, et al. "Robots with language: Multi-label visual recognition using NLP." Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE, 2013.
[11] Chambers, Nathanael, Shan Wang, and Dan Jurafsky. "Classifying temporal relations between events." Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2007.
[12] Chambers, Nathanael, and Daniel Jurafsky. "Unsupervised Learning of Narrative Event Chains." ACL. Vol. 94305. 2008.
[13] Wang, Josiah, Katja Markert, and Mark Everingham. "Learning models for object recognition from natural language descriptions." (2009).
[14] Bohus, Dan, Chit W. Saw, and Eric Horvitz. "Directions robot: in-the-wild experiences and lessons learned." Proceedings of the 2014 international conference on Autonomous Agents and Multi-Agent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
[15] Csurka, Gabriella, et al. "Visual categorization with bags of keypoints."Workshop on statistical learning in computer vision, ECCV. Vol. 1. No. 1-22. 2004.
[16] Leung, T., & Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision, 43(1), 29-44.
[17] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
[18] Socher, Richard, et al. "Parsing natural scenes and natural language with recursive neural networks." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.
[19] J. Deng, A. Berg, K. Li and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us" Proceedings of the 12th European Conference of Computer Vision (ECCV). 2010
[20] http://www.image-net.org/about-stats