**The Future of an Amalgamation of Natural Language Processing and Computer Vision**

**Charu.chhabra@india.nec.com**

**Technical Lead, NEC Corporation**

1. **Introduction:**

Natural Language Processing is the domain in artificial intelligence which emphasizes on analysing, interpreting and generating the languages used by humans for a better human computer interaction for numerous business requirements. Natural Language Processing (NLP) algorithms and models are great at processing digital text, but many real-world applications use documents with more complex formats. Some of the most common examples include certain utility forms, lab results, academic papers, receipts, genomic sequencing reports, signed legal agreements, clinical trial documents, application forms, invoices, and scanned documents. Although there are many pre-trained models for numerous NLP applications they mostly streamline the text level manipulation thereby neglecting layout and style information. Perhaps, this forms to be the most vital information in the other scenario of document image understanding.

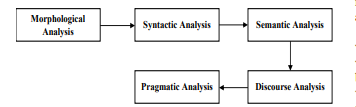
However, new advances in multi-modal learning – combining deep learning and transfer learning techniques from both NLP and computer vision – allows models to extract information from the visual documents more accurately, along with a greater degree of efficiency. NLP is an assortment to the real-world applications in numerous fields and applications such as intelligent Search engines, clinical and medical research, business processing intelligence and many more to the list.

* 1. **Working of Natural Language Processing**

Natural Language Processing expedites and encourages the machines to understand the natural language as used by humans. The language for the human computer interaction can be written or spoken, and thereby uses artificial intelligence to receive the real-life input, process it, interpret it further and finally provide the indicative meaning of the results in such a manner that the machine can easily comprehend. Similar to how humans have different natural senses like eyes to see with and ears to hear with, computers enable programme instructions to read linguistic text and microphones to gather and analyse audio. Computers have a collection of programme instructions that they can use to interpret inputs and information in a manner similar to how people use their brains to do so. This information is processed and then converted into code that can only be understood by the computer system.

* 1. **Challenges of Natural Language Processing**

Natural language processing presents researchers and scientists with a number of obstacles, most of which are related to how the process itself is always developing and changing. Precision, and occasionally its lack: Historically, communicating with computers has required employing a particular language, such as a programming language. These programming languages are very structured, precise, and free of ambiguity. However, human speech is frequently imprecise and is not always a precise mode of communication. Slang, the employment of regional dialects, and the social context of the spoken language are only a few of the complicated factors that affect the linguistic structure.



**Figure 1: Layout of the tasks involved in Natural Language Processing**

As was already mentioned, natural language processing is an iterative process that strives for perfection. Semantic analysis, for instance, continues to be a major difficulty. Another issue is how correctly such computers can understand language when it is used in an abstract way. Sarcasm is difficult for natural language processing to understand. Additionally, the emphasis that a speaker places on a particular word or syllable can alter the meaning of a statement. Speech recognition systems that use natural language algorithms risk missing the minute but crucial tonal variations in a speaker's voice. This problem is made worse by the fact that speech intonation and tone will differ depending on the accent, making it difficult for an algorithm to correctly parse. Natural language processing is hampered by the fact that human languages—and the ways in which various societies use them—are always evolving. While accepting that there are particular rules for speaking and writing a language, these might change throughout time. As the characteristics of real-world languages change, rigid computational rules and recommendations that are currently effective may become outmoded.

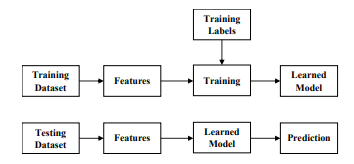
1. **Computer Vision**

Computer vision is the field of study encompassing how computer systems view, witness, and comprehend digital data imagery and video footage. Computer vision spans all of the complex tasks performed by biological vision processes. These include 'seeing' or sensing visual stimulus, comprehending exactly what has been seen and filtering this complex information into a format used for other processes. This interdisciplinary field automates the key elements of human vision systems using sensors, smart computers, and machine learning algorithms. Computer vision is the technical theory underlying artificial intelligence systems' capability to view - and understand - their surroundings.

* 1. **Applications of Computer vision**

Since computer vision may be adopted and used to create computer vision systems that can "see" and "comprehend" their environment, many instances of computer vision have been used in practical applications.

* + 1. Autonomous Vehicles: Self-driving cars employ computer vision (CV) technologies to collect data about their surroundings and
    2. Robotic Applications: Creating robots that can "see" and "comprehend" their surroundings in order to carry out their assigned jobs. Such systems merely "look" at assembly components as they move along the production line in order to inspect them for flaws and tolerance limits.
    3. Applications that search through product picture databases, recognise specific objects inside digital images, and extract data from photographs employ the CV data vision theory.

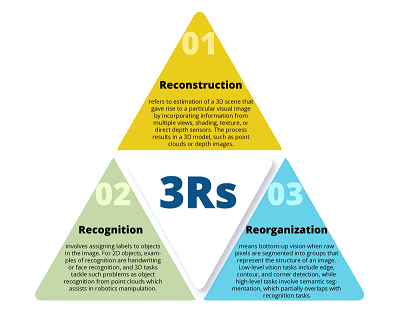


**Figure 2: Layout of the activities in Computer vision**

1. **Amalgamation of Computer Vision and Natural Language Processing**

The combination of natural language processing and computer vision involves three key interrelated processes: recognition, reconstruction, and reorganization.

* 1. Recognition: Objects in the image are given digital labels as part of the recognition process. For 2D objects, identification examples include handwriting or facial recognition, and 3D assignments deal with problems like moving object recognition that aid in autonomous robotic manipulation.
  2. Reconstruction: This process refers to 3D scene rendering given inputs from particular visual images by incorporating multiple viewpoints, digital shading, and sensory depth data. The outcome results in a 3D digital model that is then used for further processing.
  3. Reorganization: The segmentation of raw pixels into data groups that represent a predetermined configuration's design is the procedure at hand. Corner detection, edges, and contours are examples of low-level vision tasks. High-level tasks entail semantic segmentation, which might partially overlap with recognition procedures.



**Figure 3: Major components of the Integration of NLP and Computer Vision**

When compared to computer vision techniques, natural language processing tasks are thought to have a wider range of technical requirements. This diversification includes the ability to recognise varied syntax, morphology and segmentation skills, and semantics to investigate abstract meaning. Direct machine translation, dialogue interface learning, digital information extraction, and quick key summarization are examples of complex natural language processing jobs. In contrast to natural language processing, computer vision is developing more quickly. And this is largely because of the enormous interest in computer vision and the financial backing given by major tech firms like Meta and Google. Once completely integrated and combined, these two technologies can resolve numerous challenges that are present within multiple fields.

1. **Future of Integration of Computer vision and Natural Language Processing**

Customer systems can comprehend verbal or written needs in the fields of interior design, fashion design, jewellery creation, etc. and automatically translate these requirements to digital images for improved visualisation. Computer vision systems can be trained to recognise less severe human conditions and to analyse digital imagery more precisely than human medical experts. Constructing an intelligent system that 'sees' its surroundings and delivers a (recorded) spoken narrative. This outcome will be of use for visually impaired individuals. Converting sign language to speech or text to help hearing-impaired people and ensure their better integration into society. Making a system which sees the surrounding and gives a spoken description of the same can be used by blind people. The ability to sense the environment and analyse the data it gathers is provided by computer vision. Similar to how humans do it, NLP permits the interpretation of spoken or written language as well as the ability to choose the right combination of words to convey a specific message. Computers are perfectly adapted to repeatedly identify objects, recognise patterns, and report back what they perceive since they are perfect at doing these tasks.

* 1. **Medical Images in Healthcare:**

Teams of healthcare professionals frequently use CT, PET, MRI, and X-ray images to diagnose patients and choose the most appropriate treatments. In order to effectively diagnose cancer patients and create individualised treatment strategies, radiation oncologists can now use computer vision to analyse such imagery. Computer vision and NLP could be a helpful tool for performing an initial analysis of imagery and assisting doctors in diagnosing a patient, as well as for preparing an initial report of its findings and saving valuable time in the treatment process. This is because there is a shortage of qualified medical personnel.

* 1. **Visualizing creative briefs and customer requirements**

The final product might be plagued with mistakes and unmet expectations in any field requiring the design and manufacture of a physical product or space. Converting a customer's project needs into a visual could help reduce the first phases of a project, saving time and money, rather than relying only on examples of prior work that the client loves.

* 1. **Making up to bridge the communications**

The capacity to empower persons who are blind, deaf, hard of hearing, or visually challenged is arguably the most effective example of merging computer vision and NLP.

1. **Conclusion:**

Without a doubt, artificial intelligence has thus far had a significant impact on our daily lives. In some cases, we use this technology without even being aware of it in our daily lives. More than we realise, computer vision and natural language processing have a significant impact on our daily lives. Computer vision and natural language processing are fields that are constantly changing. Compared to computer vision, NLP tasks are more varied and cover everything from syntax, which includes morphology and compositionality, to semantics, which is the study of meaning and focuses on the relationships between words, phrases, sentences, and discourses, to pragmatics, which is the examination of nuances in meaning at the level of natural communication. Machine translation, dialogue interfaces, knowledge extraction, and summarization are some of the most challenging NLP activities.

In a top-down intentional method, where academics came up with a set of rules, the integration of vision and language was not proceeding smoothly. Since certain pioneers recognised some relatively specialised and narrow problems, tried a number of solutions, and found a good answer, integrated methodologies were more or less created bottom-up. The new trajectory began with an awareness that the majority of modern files are multimedia and contain associated images, videos, and writings in natural languages. For instance, a typical news item includes journalistic writing as well as a photograph pertinent to the story's subject matter. A reporter or a picture of the location where the news event occurred may also be included in a clip video. The basis for suitable and unambiguous communication is provided by the two sets of information provided by language and visual data, which are merged into a single story.