Object Recognition to Content Based Image Retrieval: A Study of the Developments and Applications of Computer Vision

Udula Mangalika
Department of Mechanical Engineering, University of Peradeniya, Sri Lanka.
ludula@aol.com

Correspondence should be addressed to Udula Mangalika : ludula@aol.com.

Article Info
Journal of Computing and Natural Science (http://anapub.co.ke/journals/jcns/jcns.html)
Doi: https://doi.org/10.53759/181X/JCNS202404005
Received 03 March 2023; Revised from 25 June 2023; Accepted 08 August 2023.
Available online 05 January 2024.

Abstract – Natural Language Processing (NLP) and Computer Vision (CV) are interconnected fields within the domain of Artificial Intelligence (AI). CV is tasked with the process of engaging with computer systems to effectively interpret and recognize visual data, while NLP is responsible for comprehending and processing the human voice. The two fields have practical applicability in various tasks such as image description generation, object recognition, and question-based answering after a visual input. Deep learning algorithms such as word input are typically employed in enhancing the performance of Content-Based Image Processing (CBIR) techniques. Generally, NLP and CV play a vital role in enhancing computer comprehension and engagements with both visual and written information. This paper seeks to review various major elements of computer vision, such as CBIR, visual effects, image documentation, video documentation, visual learning, and inquiry to explore various databases, techniques, and methods employed in this field. The authors focus on the challenges and progress in each area and offer new strategies for improving the performance of CV systems.

Keywords – Content-Based Image Retrieval, Computer Vision, Human Object Interaction**,** Natural Language Processing, Artificial Intelligence.

I. INTRODUCTION

Computer vision (CV) is an important branch of artificial intelligence (AI) that concentrates on smart computers through engineering and scientific methods. In that regard, CV researchers are dedicated to advancing computers that can see improved visual information. In this paper, the term "vision" refers to the ability to derive data from an image for the purpose of solving a particular task or understanding a process in a limited way or more generally. CV systems find use in several industrial and scientific fields, including areas such as human-machine interfaces, control of robots and autonomous vehicles, object modeling, organization of image databases, and video surveillance.

The purpose of these applications is to partly replace the human observer in evaluating visual information, including identifying which visual occurrences are relevant to resolving a certain problem. An image may be captured in its whole or in part by a CV software, which then creates a simplified representation called a model. Automated vehicles use the model to change their speed or direction of travel, robots use it to plan their next move, fire alarms use it for sound, the internet uses it to find pictures of maple leaves, brain tumors use it to describe their framework, and video game graphics are adjusted to fit the player's position. The CV model will be employed in microscopy within the framework of a cell biology investigation. In order to define the properties of a biological process, this involves replacing a cell biologist's manual inspection of photographs with a computer vision system.

Different forms of communication, including facial expressions, written texts, sign languages, and bodily gestures, are used to transmit meaning between persons [2]. Language distinguishes itself from other kinds of communication by its ability to liaise an infinite range of definitions through the usage of compositionality and syntax. The eventual importance must eventually be linked to an individual's worldview. The symbol grounding problem refers to the phenomena being described. Language without of perception would only exist as an abstract notion; on the other hand, perception without language is confined to rudimentary conditioned responses. The acquisition of knowledge about the world mostly relies on visual information for humans. The visual processing area of the brain accounts for about 29% of total brain volume [3]. A current point of contention is the degree to which language plays a direct role in the visual process. However, while developing AI, using certain linguistic components enables clarity and enhances efficient communication between people and robots. In order to demonstrate a correlation between vision and language, let us first by examining some of the fundamental goals in both vision and language.

Language refers to a collection of symbols or rules designed to produce or transmit information. The field of NLP is majorly designed to be employed by users who have minimal understanding of machine terminologies or lack enough skillset or accessibility to understand new language. The field is a specialty, which falls under the domain of linguistics and AI. The system is project to simplify the tasks performed by humans and address communication trends with computers using natural language. Natural language can be classified into Natural Language Processing (NLP), which concentrates on the study of languages, and Natural Language Generation (NLG), which is responsible for reading written languages. Phonology alludes to the learning of sounds in different languages, whereas Vocabulary relates to the development of vocabularies. Syntax relates to the structure of different sentences; Semantics concentrates on defining linguistic meanings; and Pragmatics relates to the comprehension of languages and employing it in context. The 20th century pioneer of linguistics, Noam Chomsky, transformed the study of grammar based on theoretical languages with significant advancements. NLG defines a system for defining significant and coherent sentences, and classifies it using internal representations.

This paper provides a detailed account of various CV-based fields, such as CBIR, visual questions, video captioning, image captioning, and object understanding. In this research, the developments and challenges of various fields are provided, examining the application of deep learning approaches to enhance efficiency. The paper highlights the significance of NLP in the field of CV, and the integration of visual elements and linguistics. The principle aim of this research is to provide a rationale for applying CV and its application in various fields. The rest of the paper is arranged as follows: The second section defines the interconnection between CV and NLP. The section explores various NLP and CV application. The third section focusses on the relation between visuality and language in the field of multi-media. This section covers basic topics such as visual feature identification, CBIR, visual querying, and visual identification. Lastly, the fourth section presents final remarks regarding the research.

II. THE CONNECTION BETWEEN CV AND NLP

CV Tasks and NLP

Computer vision (CV) allows computer systems to effectively comprehend and evaluate visual data. These systems employ sophisticated visual and deep learning models to accurately identify and categorize objects, and provide instant feedback. Within the domain of artificial intelligence (AI), CV depends on automated models, which can review visual data, such as images or films, in a manner that mimics the perception of humans. He teaches computers to analyze images, and they are understood at a granular level by examining each pixel individually in computer vision. This is the basic foundation of the CV project. Computers will acquire visual information, process it, and evaluate results with advanced software programs, with special emphasis on the technical side. Consideration of the 3Rs, i.e., restructuring, accreditation, and repetition, includes CV responsibilities.

To reconstruct is to determine the original, three-dimensional (3D) environment from which a certain image or rendering was created. Integrating data from many views, using depth sensors, appearance, or shading are just a few of the possible approaches to completing the job. A three-dimensional model, created during the reconstruction process, might appear as depth images or point clouds. Reconstruction tasks include several activities such as shape from shading, scene reconstruction, and Structure from Motion. Recognition comprises several issues in both 2D and 3D domains. In the 2D domain, it involves tasks like recognizing handwritten writing, identifying faces, sceneries, and objects. In the 3D domain, it involves recognizing 3D objects from point clouds, which is particularly useful for robot operations. Recognition entails the allocation of labels to the objects shown in the image.

The use of prepositions is common when describing the relationships between things or between an item and its environment. Adverbs may be thought of as the links between an item and an action in terms of time. Changes to a core set of components that may be considered the building blocks of forms, textures, colors, areas, and movements are among the challenges of reorganization. The fundamental elements amalgamate to create higher degrees of visual cognition: These words frequently lack a specific subject or context that may be expressed, however they are fundamental for learning new vocabulary since they tacitly transmit the attributes of things or settings. Reconstruction sometimes integrates real-world physics and 3D geometry, offering more comprehensive item or scene attributes in comparison to rearrangement jobs. Real-time robots are using reconstruction techniques with high precision by utilizing the acquired depth dimension.

NLP Tasks and CV

Three design categories may be used to group machine translation systems: direct, transfer-based, and interlingua-based systems. **Fig 1** shows the famous "Vauquois Triangle" (VT), a common way to represent these three tactics. At the topmost level of the triangle is Interlingua machine translation, in the very center is transfer-based translation, and at the very bottom is direct translation, representing the three levels of hierarchical representation. As we progress from the direct method to the transfer strategy, and eventually to the Interlingua approach, the amount of analysis required expands. Moreover, it offers understanding of the decreasing amount of transferable knowledge required as we go up the triangle. At the literal level, a substantial quantity of data is sent, with every word imparting knowledge. Shifting guidelines are only required for parse trees and thematic objectives at the transfer level. Ultimately, at the interlingua level, a language-neutral conceptual framework that represents both the source and destination texts is encoded.

The first technique, which has largely been discarded by machine translation researchers, entails incorporating all the specific information for a given language pair into a single translation direction. When using the direct technique of

translation, it is essential to keep the link between words as exact as possible. The second choice uses post-disambiguation information provided by internal syntactic representations. As a first step, the source language-specific representation of the original writing is converted. The next step is to transform this representation into a language-specific representation, and then utilize it to create text in the aimed language. The third perspective makes use of Interlingua, a fully independent language, which represents textual sources, which are reviewed; and target language texts, which will be produced. An existing machine translation model may have its translation capacity increased to integrate more language pairs using the most applicable method.

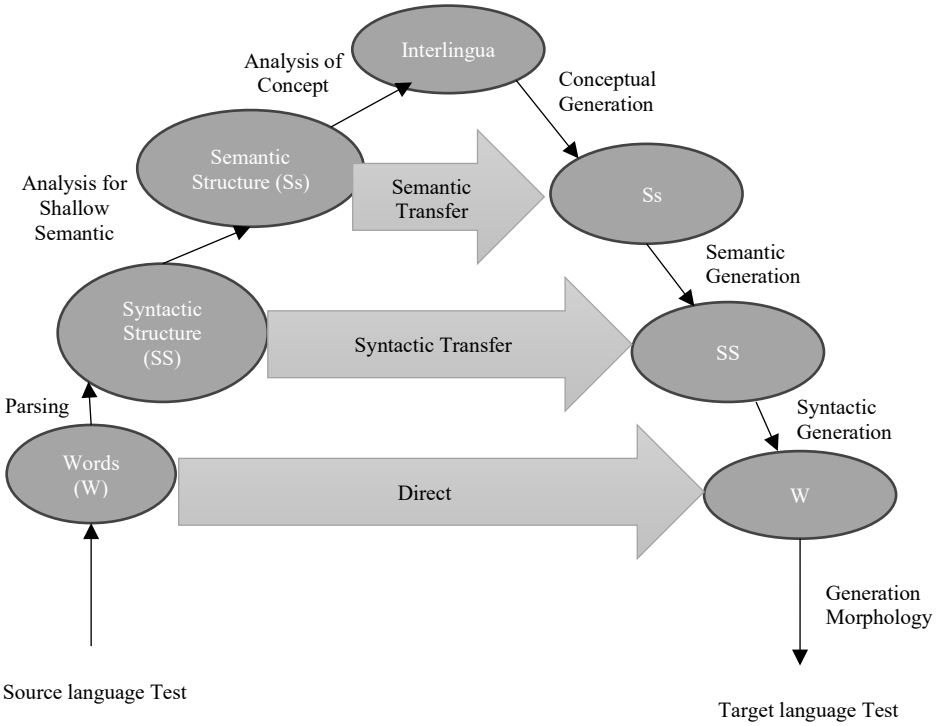

Fig 1. VT Methodologies

Vauquois triangle refers to a framework employed in machine translations to categorize NLP activities into various language levels, such as semantics, syntax, and pragmatics. The capacity to accurately communicate is improved by this form of categorization [14]. Morphology, a review of work formats and compositionality, the process by which minor linguistic units, such as works are integrated to create larger units, such as phrases, or sentences, are all a major segment of syntax. Research in the field of semantics concentrates on the connection between expressions, sentences, phrases, and words in an aim to effectively determine their meaning. The major concern in academics relates to the manner in which contextual instances affect the meaning in pragmatics.

To completely understand the intended purpose of the speaker in an ironic statement, it is fundamental to understand the context. The abundance of ambiguity in correct interpretation of languages is a fundamental barrier for a smart system to achieve full language comprehension. Natural Language Processing (NLP) integrates various complex activities such as dialogue summarization, interface, question answering, data extraction, machine translation, parsing, and data extraction. The translation process to various languages normally results to gradual loss of the original value and intent of phrases and words.

There is a substantial gap that must be bridged when going from a simple image representation (like pixels or outlines) to a more complex one (like employing language labels or words). The idea of linking visual information, such phrases, or words, with language data is called “bridging the meaning gap,” as described by [5]. The act of assigning a label to an image that contains an object is known as item recognition. Scene recognition is the process of labeling an image's backdrop with descriptive information. The term “semantic segmentation” describes the method of grouping pixels according to their semantic content. Visual processing might benefit from understanding word correlations, which could help clarify various visual structures. A knife, for instance, is more appropriate for slicing cucumbers than a chair in the kitchen since the former's interplay is based on a concrete notion within the physical realm, while the latter is more at home in the abstract.

Grouping the present language and vision landscape

The development of techniques that combine visual perception with language understanding did not follow a systematic and hierarchical methodology, where scholars established a set of rules that determined the components of the study. The approaches were typically developed using a bottom-up strategy, whereby pioneers determine specific difficulties in distinct regions, experimented with several possible solutions, and recorded good results. To provide a framework for this

circumstance, we have designed a matrix where the rows represent the columns and vision corresponds to thinking and language.

III. MULTI-MEDIA VISION AND LANGUAGE

Visual Feature Description

Attribute-Based Vision.

Visual attributes alludes to particular semantic aspects that may be pre-determined to describe various objects, including characteristics such as shape, color, and substance. Jiang et al. [6] compiled an AWA dataset of 37,322 photos representing 50 distinct animal types. In addition, they provided an attribute distribution consisting of 85 dimensions for each category. Yaghoubi et al. [7] developed a PETA dataset with 19,000 pictures depicting 8,705 pedestrians. In addition, they included 61 binary attributes and four multi-category variables for each wayfarer. A dataset named aPY was created by Ashrafi, Shokouhi, and Ayatollahi [8] by tagging 12,679 specimens of Pascal VOC. Every test in this dataset is given a 64-dimensional attribute description. In their study, the authors in reference [9] provided a comprehensive analysis of 204-D ascribe angles for 29 different kinds of tests in the COCO dataset. However, a mere 18,073 examples had labels including over 10 positive features.

Attribute learning is widely employed in different CV areas such as semantic segmentation, zero-shot learning, face analysis, fine-grained recognition, and human reidentification. Kulis and Grauman [10] introduced a methodology that included all category labels into the attribute vectors space to achieve precise identification of unknown things. In their study, Han et al. [11] recommended an attribute-aware attention model to enhance the precision of fine-grained picture categorization. This paradigm seeks to include both the depiction of global categories in the world and the depiction of specific attributes at a local level. The research conducted in [12] entailed simultaneously obtaining pedestrian characteristics and identification labels utilizing a common backbone. A Convolutional Neural Network (CNN) is a kind of artificial NN that is particularly effective at analyzing visual data. This approach integrates attributes with subtle visual cues to enhance the precision of reidentification. Vu and Huang [13] devised a multitask convolutional neural network (CNN) that simultaneously acquired knowledge on the center position, size, displacement, and semantic advancements. This CNN was purposefully designed for the sole objective of detecting pedestrians. A comprehensive semantic segmentation technique was created in [14], which replaced the single category “person” with nine different attribute categories.

However, most existing researching algorithms are restricted by the capacity of human-defined ascribes to effectively represent the required features. Consequently, some scholars have proposed the acquisition of latent features to get extra distinctive characteristics. Gao [15] introduced the concept of a semi latent attribute space, which integrates user-latent and prescribed features under a unified framework. In addition, they provided an ascendable possibility topic model specifically tailored to learn the multichannel semi latent features. Ye and Guo [16] proposed a novel wordbook learning model that categorizes the wordbook capacity into three components, each reflecting latent background, latent discriminative, and semantic properties. Wang et al. [17] developed a comprehensive network that can acquire separate enhanced spatial semantic representations for both latent and user-defined characteristics. In their study, Wang, Zhang, Ji, Pang, and Ma [18] proposed a technique for acquiring a semantic dictionary that utilizes concealed visual characteristics and concurrently aligns the visual-semantic domains. This research presents a novel method for object segmentation and identification called self-supervised ascribe studying methodology. Our technique differs from the previously described approaches by not requiring extra attribute annotations. Instead, it focuses on learning latent characteristics that enhance cross-class transfer learning of human-defined attributes.

HOI Detection and Prediction.

Zhang, Rojas, Liu, and Guan [19] conducted groundbreaking research on visual semantic role labeling in the context of existing HOI detection systems (refer to **Fig 2**). The objective of this work is to ascertain the exact position of the agent (human) and object, as well as discerning the engagements between them. The research carried out by Gao, Zou, and Huang [20] introduces a human-focused approach called InteractNet. This approach improves the architecture of Faster R-CNN by including a summation branch that specifically learns the density map for target areas. Vahora and Chauhan [21] propose using a graph CNN to address the task of handling human-object interactions (HOI) as an optimization problem that involves graph structures. A multi-stream network is created by Liu, Mu, and Huang [22] using the branch representing paired interactions and the area of interest for HOI. The multi-stream design takes the real image, the Feature Pyramid Network (FPN), and the anticipated bounding boxes (BB) from the pre-trained detectors as inputs.

The visual properties retrieved from the backbone network are used to identify individuals and objects in a multi-stream architecture. These attributes are used to generate trust estimations for the BB of the detected individuals and items. The paired stream captures the geographical relationship between the object and person by merging the two bounding boxes (object and human) into a unified entity. Following research have further developed the previously described multi-stream system by including new components such as posture information, deep contextual attention, and instance-centric attention oriented on context-aware feature of appearance on individual scores. The scores are derived from distinct streams that represent an item, a person, and paired interactions. These values are then integrated using a late fusion technique to recognize interactions. According to Wang, Huang, and Zhang [23], relying just on appearance data is insufficient for accurately representing complex human-object interactions, which makes the late fusion approach less than optimum. Furthermore, the pairwise stream merges the individual person and item boxes to create a reference box that is used to

generate a binary image representation. Nevertheless, this methodology might provide inaccurate forecasts due to its dependence on restricted geographical data.

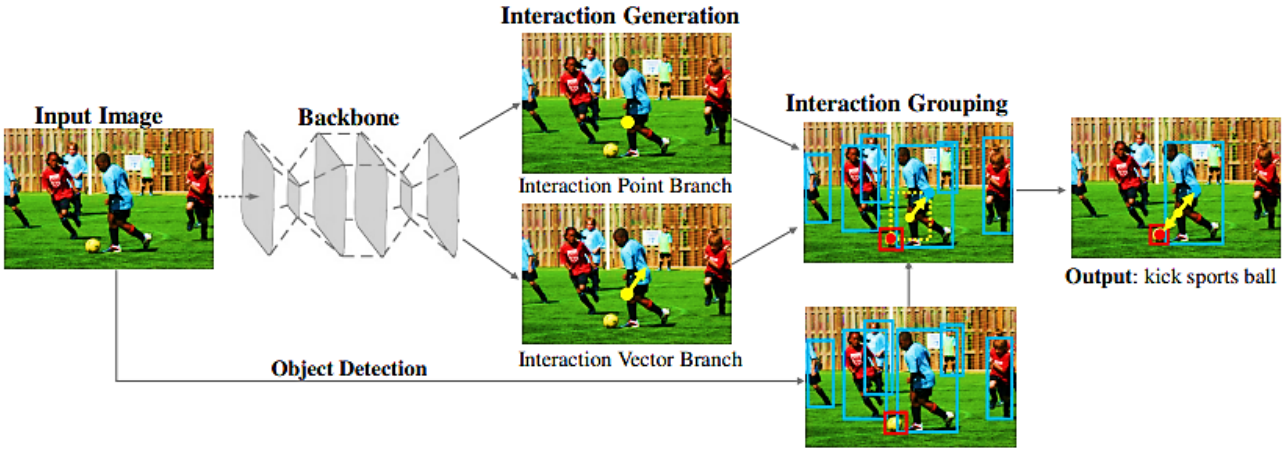

Fig 2. The Structure of the Suggested HOI Detection Framework

Consistent with previous research [24, 25], we use a Feature Pyramid Network (FPN) to provide predictions for bounding boxes of objects and humans. The stage of interaction forecast is made up of three consecutive steps: interaction grouping, interaction generating, and feature extraction. The production of interaction involves two independent branches that produce an interaction vector and an interaction point. The final projections for HOI—human, action, and object—are then provided by the interaction grouping module upon input of the interaction vector and point as well as the bounding-box predictions [26] for people and objects.

Understanding actions is a necessary component of predicting object properties. Several distinct verbs may be used to describe any human activity, depending on the viewpoint of the speaker. All verbs, on the other hand, might involve a particular set of acts. Verbs may operate similarly to how ascribes function for objects. Transitive verbs often need concrete things and human subjects to function as their predicated complements. (This is discrete from character-based behavior detection in humans, which is similar to anticipating object properties.) Verbs, objects, and human attitudes must all be identified in order to identify activities. Discriminative classifiers, such Support Vector Machines (SVMs), are mostly used in recognition. Nevertheless, in this specific situation, it is crucial to consider the disparity in class allocation and the correlation across many categories. Verbs include a wide array of visibly distinct “doing acts”, such as using different tools, postures, and methods when comparing the operation of felling a tree to slicing a cucumber.

Prominent datasets used for forecasting human-object interactions (HOI) include Stanford 40 action, Humans Interacting with Common Objects (HICO) [27], and Trento Universal HOI (TUHOI) [28]. Before TUHOI and HICO, the datasets that were used were often of a somewhat restricted magnitude. Kumar, Lahiri, and Ojha [29] contend that only forecasting HOI without integrating semantic role comprehension is inadequate, since this forecast fails to provide comprehensive and nuanced data in the resultant output. The primary objective of the V-COCO dataset is to demonstrate persons engaging in several tasks concurrently, while also assigning specific duties to each object involved in each activity. In contrast to typical HOI forecast algorithms, V-COCO argues for the use of object locators rather than differentiative classifiers. This text provides an initial overview of satisfactory function stamping, a well acknowledged topic in the field of NLP, particularly its connection to CV. Frame Net and Verb Net ontologies are linguistic resources that function as a collective knowledge repository for NLP by offering general knowledge.

Similarly, Zhong et al. [30] provide a baseline for structured prediction using the imSitu dataset, which is a Conditional Random Field (CRF) constructed on a CNN. As can be shown in their research, the baseline performs better than more conventional injustice classifiers such as SVMs. Using this technology to precisely assign semantic roles to things in images linked to actions would improve the creation of intelligent conversational bots.

Visual Description

Image Captioning.

Image Captioning (IC) is the act of verbally describing the visual content shown in a photograph. This problem is located at the intersection of CV and NLP. Most IC methods have an encoder-decoder architecture, where an input image is converted into an intermediary delineation that apprehends the inherent data of the image. This interpose delineation is then converted into a logical and instructive textual sequence. The most often used benchmarks are COCO and nocaps. Models are often assessed using the CIDER or BLEU measure. A phrase's structural arrangement paints a clearer picture of an idea than a disorganized collection of words. A story's textual narration is often accompanied with illustrations. For instance, a sports headline picture can show the crucial moment of the match, but the writing that goes with it will offer a detailed explanation. In order for words to have a certain meaning, their interpretations need to be easily understood.

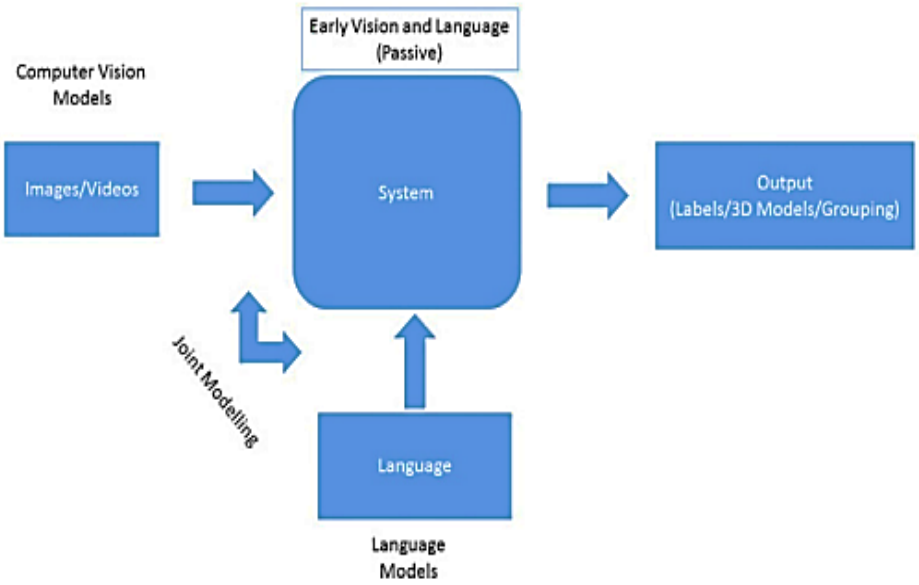

Fig 3. The Early Vision and Language Framework

Shen et al. [31] conducted a research that investigates the relationship between BLEU, ROUGE, and METEOR metrics and human assessments. The study reveals a weak association, particularly for BLEU. METEOR has the powerful connection, however it still falls somewhat short of true human judgments. Another approach to evaluate is using human review services such as CrowdFlower or Amazon Mechanical Turks. These models utilize numerous criteria like as relevancy, completeness, readability, correctness, and human likeness. Nezami and colleagues [32] provide a further examination of picture captioning systems. In photo captioning, the majority of systems usually focus on including language data as an extra layer or simultaneously combining vision and language via a well-constructed loss algorithm or function, as seen in **Fig 3.** These systems use sophisticated algorithms to examine intricate input from numerous sources and provide structured output, distinguishing them from traditional systems.

Video Captioning.

This concept is extensively used in contemporary community across several applications to accommodate those with disabilities, including those who have visual impairments. Thanks to recent advancements in object recognition and NLP, there has been a notable surge in the incorporation of these functions into widely used apps. An example of this fusion occurs when picture captions are created by the system based on an input image, providing a brief description of the contents of an image. The initial emphasis of this fusion was on photos, which was then expanded to include videos, including some adjustments to the preexisting methodologies.

A plethora of contributions from people worldwide have been attained in this topic. Therefore, it was imperative to compile, analyze, and assess all the discoveries and present them in a comprehensive study, as shown in [33]. An assessment was carried out to evaluate several video captioning algorithms on a particular dataset, using various criteria often used for image and video analysis. The assessment done by Graham et al. [34] specifically examined the approaches employed from 2015 to 2019, analyzing each year separately. The predominant dataset and assessment approach are often shown graphically by bar graphs and scatter plots, with each year displaying the associated evaluation parameter. Notwithstanding extensive examination and investigation carried out on video captioning, our survey uncovers many concerns.

A video may be described as a coherent and structured compilation of language that tells a certain story. One may gather prior knowledge from extensive online data sets to shape the model and discourage unlikely pairings of items, activities, and humans. By incorporating motion tracking with event semantics, the accuracy of understanding activities in a video within a certain context may be greatly improved. This approach allows for the identification of actions, forces, and dynamics associated with the activities. Borghi and Riggio [35] were the first to present a method similar to the phrase tracker. By combining natural language semantics with three crucial cognitive visual tasks—event recognition, object identification, and tracking—a link is built between the languages` structural arrangement and the hierarchical structure of video events. An integrated cost function that assimilates the attention technique to categorize the most essential event allows all three tasks to be executed concurrently, emerging in the creation of the optimal descriptions for activity identification in phrases. The sentence tracker make use of lexical linguistics to collect the specific data on the agent, manner, recipient, and location of an action.

An object is characterized as a noun phrase, but the seen behavior is indicated as a verb. Object attributes are defined via the use of descriptive words, while the relationships between things in space are conveyed through the use of prepositions. Prepositions and adverbs express different characteristics of an event.

The technique uses a pre-existing vocabulary and, using a set of a predetermined grammar principles, produces a phrase. It is possible to split the sentence tracker into two parts. The First Subsystem (FS) is divided into three distinct stages. At first, a setup is used for object detection with an emphasis on getting a good recall rate. In addition, forward projection enhances the tracking process by increasing precision. In addition, the optical flow-matching output is achieved by combining the Viterbi approach with dynamic computing to choose the optimal collection of detections. Employing the Viterbi algorithm for computing, the second subsystem (SS) uses Hidden Markov Models (HMMs) to detect events. Due to their shared reliance on HMMs, the final target functions obtained from the FS and SS may be mixed. Three distinct cases were used to test the sentence tracker: Topics covered include video retrieval, sentence production, and directed attention to complete sentences.

Visual Question Answering

There has been a lot of discussion in the disciplines of CV and NLP about Visual Question Answering (VQA). Not long ago, this work was acknowledged as an AI-complete challenge, meaning it might theoretically replace the visual Turing test. A computer's ability to correctly answer a query about a picture using natural language is essential to this discussion. The impressive feats it has accomplished in speech, text, and vision technologies are what draw many deep learning experts to it. Since VQA generates textual descriptions of images, it shares certain similarities with picture captioning. Since it requires an in-depth knowledge of images that cannot be acquired only via general explanations, accurately responding to requests with visual information is a challenging undertaking. Since the correct response must pinpoint the specific positions of objects, their surrounding space, and the spatial connections between them, the first approach, often called the traditional " to determine the location just by gazing," is still in use. Additionally, the contextual data should have the capacity to respond to inquiries on the precise timing, purpose, and approach. The visual aspect of a scene might differ considerably between the seasons of summer and winter. It is recommended to use a shovel to clear the collected snow piles in front of the house in this specific situation.

The main challenge in Visual Question Answering (VQA) is in devising system mechanisms and a semantic representation that can efficiently extract data and perform logical reasoning on that information. The understanding of the significance relies on the specific qualities of the evidence. For a static image, it is often sufficient to include information about the connections between items and scenes. The system is required to generate sentence descriptions, as stated by Roy [36]. If the data is shown as a series of images or videos, the system must account for the passage of time in order to accurately portray the progression of actions, events, or any other kind of dynamic change. Like the vQA dataset given by Antol et al. [38], the question-answering dataset proposed by Kafle and Kanan [37] is comparable. A memory network is used by the proposed system to record the relationships between the narrative, the question, and the response while the question is being answered. Another important factor to think about is how the system plans to get the necessary data. Yu et al. [39] developed a dataset called visual madlibs for Visual Question Answering (VQA), which calls for a system that can answer both multiple-choice and complete the blank questions.

An organized framework for visual inquiry is provided by visual madlibs. This assists the system in identifying the best strategy, whether attribute prediction for inquiries about an object's attributes (type 6) or saliency detection for answering intriguing questions about the picture (type 3). Using joint-embeddings or deep-learning algorithms is a straightforward and successful strategy, according to the dataset developers. Including a classification of Visual Questions (VQ), Visual7w is an additional Visual Question Answering (VQA) dataset. Where, when, how, who, and why are the seven types of inquiries that Visual7w classifies (7w). These categories are presented in a multiple-choice manner. Visual7w is a branch of Visual Genome, an extensive dataset that seeks to build associations between images and WordNet. The information it includes is diverse and covers several aspects, including object instances, captions, area descriptions, question responses, relations, and properties. The collection contains a vast number of instances for each of these contextual components.

Current approaches (see **Fig 4)** may use a cognitively inspired technique to analyze the architecture of visual and verbal modalities in order to solve visual question answering (vQA). Built on top of an active vision design, the systems may choose their own next sample to study. The attention mechanism (AM) governs the manner in which the output might provide feedback that influences the underlying features inside the system. The REINFORCE algorithm, introduced by Williams [40], is a kind of reinforcement learning system that may naturally reach this conclusion by using an exploration-exploitation framework, as detailed by [41]. The agent's architecture include a memory module and may be trained using the usual backpropagation method. Moreover, the system must effectively tackle the concurrent incorporation of visual and verbal components as a difficulty in linking symbols and messages. In essence, the system has to create a distinct and conclusive link between the two modalities.

Content-based Image Retrieval

Language, specifically query concepts or strings, is crucial to the multimedia subfield known as CBIR. As shown in **Fig 5,** the main parts of CBIR consist of mandatory phases and possible extra stages. Phase one of CBIR requires the user to submit the query image. All the images in the databank, including the query image, will go through the same applicable operations in the sequence that is defined. These operations are often known as online processing and are typically applied to the query image in real-time as the user submits it. Prior to query submission, offline procedures may be performed on dataset images.

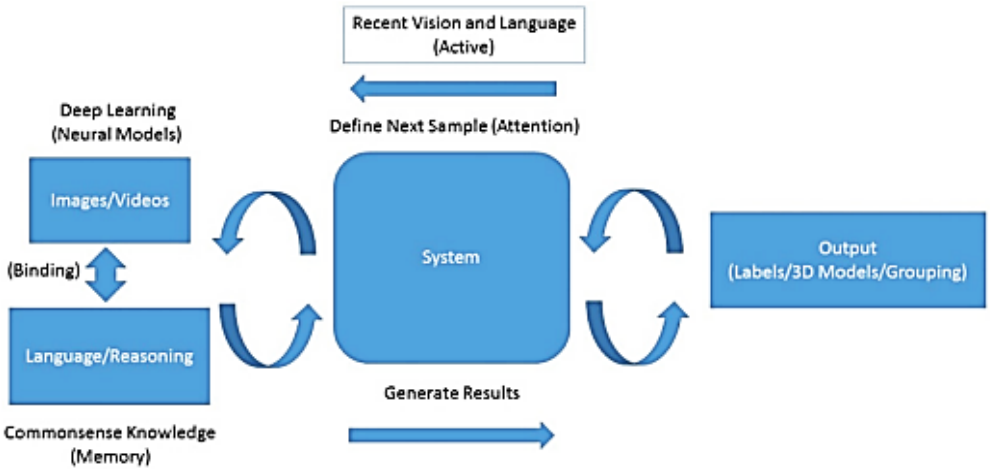

Fig 4. The Modern Language and Vision Model Representing Cognitive Processes.

Resizing, scaling, reduction of noise, and categorization are some of the possible activities that may be included in an optional preprocessing step depending on the framework's architecture. Feature extraction, after the optional stage, is often thought regarded as the most important stage. Transforming an abstract concept into a numerical one is what this stage is all about. Local descriptors or low-level attributes including shape, color, texture, and spatial information are two possible formats for the retrieved features. As an additional preprocessing step following feature extraction, normalization or classification is at your discretion. Finally, all the photographs in the collection are compared to the query image's properties to determine which ones are most relevant. By letting users actively participate in deciding whether the shown images are relevant or irrelevant, relevance feedback is an additional procedure that improves the results. Some approaches have been recommended to provide pertinent feedback with the aim of enhancing the performance of CBIR models.

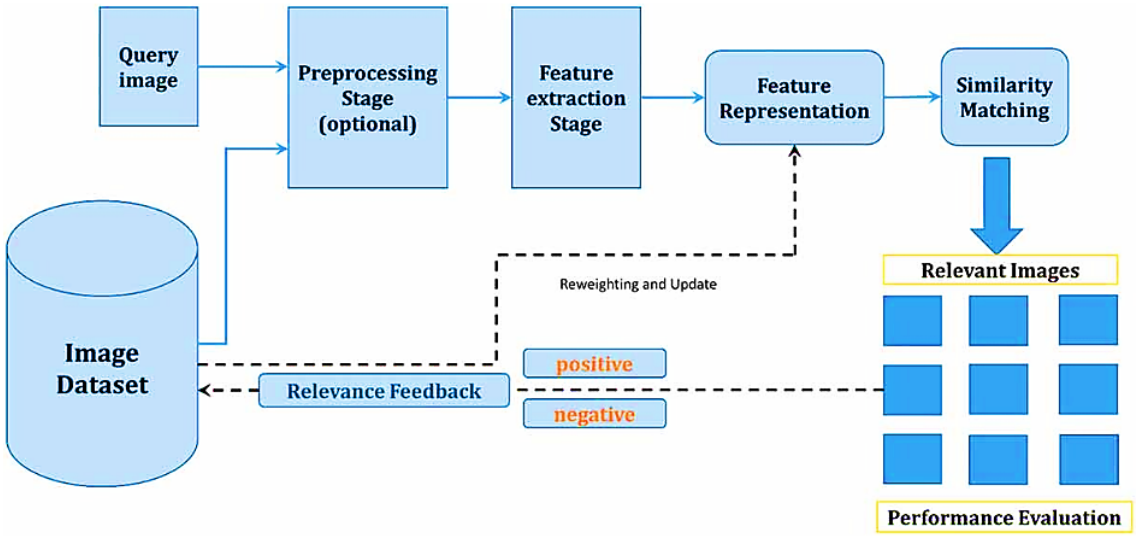

Fig 5. Overview of the CBIR Infrastructure

Historically, images have been classified according to their fundamental visual attributes such as texture, form, and color. Contemporary CBIR systems strive to assign a word to a specific part of an image, similar to semantic segmentation. The goal is to generate keyword tags that people can really understand. CBIR systems use visual features to represent an image for the objective of comprehending it, whereas keywords are used to classify images for the particular objective of image retrieval. Since visual traits may modify themselves to fit each target domain, they are an ideal candidate for the CBIR intermediate layer. The reason for this distinction is that CBIR has a lower level of noise as compared to other fields of application, such robotics.

Different CBIR [42] approaches investigate the relationship between text and image content using different models. There are a number of different approaches to these models. Some of them include: word co-occurrence models in image grids, automated translation models that turn blobs of images into words, statistical models that do the same, topic models that look at both words and blobs, the Cross-media Relevance Model [43] which delves into the Continuous-space Relevance Model [44], and a combined distribution of words and image blobs, which prioritizes semantic features over visual ones like

texture, form, and color. According to Zheng, Zhang, and Chen [45], it is proposed to enhance supervised Latent Dirichlet Allocation by including the hidden spaces of picture annotation and image classification. This is based on the assumption that there is a potential relationship between the labels and tags. By extending Canonical Correlation Analysis (CCA) with a kernel, Melzer, Reiter, and Bischof [46] are able to bring data from the visual semantics, tags, and attributes domains into a shared subspace. Multimodal retrieval is made possible by this.

Image tagging accuracy might be enhanced by using deep learning techniques and word embedding, as shown by Gong et al. [47]. The photos and their tags undergo intermediary feature adjustments to accomplish this. The following datasets are often used by CBIR: NUS-WIDE, Corel5k, MIRFlickr08, ESP Game, PASCAL-VOC, UIUC, IAPR TC-12 and Labelme. Müller, Michoux, Bandon, and Geissbühler [48] provide a comprehensive analysis and comparison of current methodologies and emerging challenges in the field of CBIR. Another facet of attention in CBIR is the retrieval of films based on certain incidents or events. The responsibility is closely connected to the process of identifying activities, since it involves considering the time element present in a video.

One way to portray text visually is by word embedding. It is possible to find word similarities using text vector representations. Improvements in embedding methods have made deep learning more useful in natural language processing. When it comes to picture categorization, the Word2Vec Skip-Gram model is a dead ringer for the CNN model's design. In addition, it adeptly analyzes musical semantic linkages and makes good use of computing resources via the use of parallelization in shared and distributed memory situations. To find out how similar the two words are, we apply the Word2Vec and LDA domain models. By comparing word similarities, a semantic network may be constructed. By grouping words into several communities, each of which represents a unique idea, community recognition algorithms can automatically extract ideas from text. Biometric surveillance systems that utilize GloVe integration with the BiLSTM framework considerably outperform those that do not, according to research by Pimpalkar and Raj [49]. The importance of word embedding is shown in **Table 1**, which presents a thorough examination of its performance evaluation, data origin, datasets, and areas of application.

Table 1: A Review of Work Embedding Significance

Embedded Approach	Dataset Name	Application Areas	Architecture	Efficiency
Word2Vec	CIFAR-100, ImageNet	CNN representation	VGG16, Inception-v3, ResNet50, CNN	Effectively signified the word
Fasttext	One Billion Words benchmark dataset Small text8 dataset from Wikipedia	Parallelizing the algorithm	Negative and mini batching sample sharing	Multi-core architectures are scaled by Word2Vec
Word2Vec	USPTO-2 M, USPTO, WIPO-α data set	Patent domain downstream tasks	BiGRU	MAP—64%
Word2Vec, fastText	STS-dataset 2012–2016	BERT model multiple layers	SBERT-WK model	Sentences effectively represented by SBERT-WK
Word2Vec	MIDI dataset	Word2Vec's semantic correlations in music	Slices based on spatial proximit	Harmonic relationships and meaningful tonal captured by Word2Vec
BERT	Turkish texts dataset	Hyperparameters	The random search and grid search	Accuracy—93.8%, F1-score—89.7%
Word2Vec	Turkish dataset	Turkish excision of corpus relations	Method for projection optimization	Accuracy-90.75
Word2Vec, Law2Vec	World Intellectual Property Organization and Board of Veterans Appeals (BVA) cases	Legal analytics	LSTM, BiLSTM CNN	BiLSTM + CRF + Law2Vec gains F1-score of 88%
fastText, Word2Vec, Law2Vec, GloVe	TREC, RT movie review, Movie reviews, CR from	Sentiment analysis	CNN	An accuracy of 87%is achieved by the CNN + IWV

	Amazon, Stanford sentiment treebank			
Word2Vec, GloVe, BERT	Data from Cooling Plants and Central Heating Plants at the California University, Davis campus	Next alarm prediction in process plants	LSTM	81.40% accuracy is achieved by LSTM + Word2Vec
Word-position2Vec, Word2Vec, Lexicon2Vec, GloVe, Pos2Vec	American online dictionary Disaster Tweets dataset	Mirroring vector space approach	Mirror Vector Space model	An accuracy of 83.14% is achieved by the proposed model
Word2Vec	Text8 corpus, MEN dataset, and Word-Sim353 (WS)	Word2Vec Scaling on a GPU Cluster	High-level Chainer DL model	Higher results at the subword level are achieved by the proposed framework with Word2Vec
Word2Vec	Sequence-to-sequence autoencoder (SSA)	Phonetic information extraction	RNN	Increased accuracy using Word2Vec and SSA
AudioWord2Vec	IMDB and Wikipedia	word similarity, text classification, topic modeling	Probabilistic approach	F1-score76.9%, Precision-82.8%
GloVe	IMDB dataset	Word Similarity	Isotropic Iterative Quantization (IIQ), CNN	The 76.43% accuracy is achieved by the proposed IIQ approach
GloVe, fastText	WikiText dataset	Word embedding for certain tasks	The algorithm for post-processing	Better embedding with small corpora is yielded by GloVe + Post-Processing approach
GloVe, fastText, and Word2Vec (Skip-Gram model)	Datasets for categorizing sentences	Dimensionality reduction	The algorithm for post-processing	With the suggested method, a Spearman rank correlation value of 91.6 is attained.
Word2Vec, GloVe	Talk about the procedures used in each embedding	Deep learning environment	CNN, RNN	Discuss the effectiveness of DLM
GloVe	WordNet dataset	Synonymy detection	LSTM	synonymy detection is improved by Semantic similarity among noun vectors
Word2Vec, Adaptive Cross-contextual Word Embedding model, ELMo, BERT, GloVe	Dataset from AWSDL-TC3, Wikipedia corpus	Extract API information	Gaussian Latent Dirichlet Allocation (GLDA)	Web service discovery is well handled by the G LDA framework
Word2Vec	SST dataset, TREC dataset, Review data from Amazon	Dimensionality reduction	LSTM, CNN	93.48% accuracy on TREC dataset is achieved by the distillation ensemble strategy
BERT, Word2Vec, OpenAI-GPT, GloVe, ELMo, fastText	Knowledge Encyclopedia Dictionary and Assault Incident News Articles	Word extraction for Korean language	Backward mapping and skipping method	Effective performance is achieved by the proposed approach

DSWE: domain-specific word embedding, SA: sentiment analysis, IWE: impact of word embedding, TC: text classification, RS&NER: recommendation system and named entity recognition system, MTC: medical text classification, TM: topic modeling,

IV. CONCLUSIONS

Computer vision is a progressive domain inside artificial intelligence that concentrates on instructing computers to scrutinize and comprehend visual input. The core operations performed by systems that use computer vision may be succinctly

summarized as reconstruction, identification, and rearrangement. There is a strong correlation between the challenges faced in NLP and CV. Academics use many approaches and concepts to develop links between these two fields. Learning techniques have been devised to improve the precision of detailed classification of image, but their successfulness is limited by the quantity of characteristics supplied by people. Researchers have proposed that latent characteristics and hidden visual elements might potentially address this limitation by improving the process of acquiring attributes. The paper introduces a method for acquiring characteristics in the processes of object recognition and segmentation via self-supervision.

Moreover, it reviews InteractNet, a human-oriented approach used to identity and analyze interactions between humans and machines. This paper delves into the domains of visual question answering and video captioning, both of which necessitate complete comprehension of images, the capacity to extract data, and logical reasoning, and both pose significant challenges. Content-Based Image Retrieval (CBIR) models utilize different methodologies and models, such as deep learning and word embedding, to define and retrieve images such as natural language methods, thus enhancing the precision of image recognition. One critical method in the domain of NLP for identifying the similarity of words, and extracting significant concepts from texts is word embedding. Opportunities for future research and development abound in the domain of CV, mostly in feature acquisition, video description, visual question answering, and HOI detection. Enhancing efficiency and accuracy of these processes will have significant advantages in areas such as robotics, autonomous vehicles, and medicine. Word embedding approaches and CBIR models are continuing to advance significantly, which will enhance the efficiency of NLP.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

There are no competing interests.

References

[1]. M. Leo, G. Medioni, M. M. Trivedi, T. Kanade, and G. M. Farinella, “Computer vision for assistive technologies,” Computer Vision and Image Understanding, vol. 154, pp. 1–15, Jan. 2017, doi: 10.1016/j.cviu.2016.09.001.

[2]. S. C. W. Ong and S. Ranganath, “Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 873–891, Jun. 2005, doi: 10.1109/tpami.2005.112.

[3]. J. H. Kaas, “Why does the brain have so many visual areas?,” Journal of Cognitive Neuroscience, vol. 1, no. 2, pp. 121–135, Jan. 1989, doi: 10.1162/jocn.1989.1.2.121.

[4]. B. Kituku, L. Muchemi, and W. Nganga, “A review on machine translation approaches,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, no. 1, p. 182, Jan. 2016, doi: 10.11591/ijeecs.v1.i1.pp182-190.

[5]. D. K. Mishra, A. Thomas, J. Kuruvilla, P. Kalyanasundaram, K. R. Prasad, and A. Haldorai, “Design of mobile robot navigation controller using neuro-fuzzy logic system,” Computers and Electrical Engineering, vol. 101, p. 108044, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108044.

[6]. B. Jiang, W. Huang, W. Tu, and C. Yang, “An Animal Classification based on Light Convolutional Network Neural Network,” 2019 International Conference on Intelligent Computing and Its Emerging Applications (ICEA), Aug. 2019, doi: 10.1109/icea.2019.8858309.

[7]. E. Yaghoubi, D. Borza, J. C. Neves, A. Kumar, and H. Proença, “An attention-based deep learning model for multiple pedestrian attributes recognition,” Image and Vision Computing, vol. 102, p. 103981, Oct. 2020, doi: 10.1016/j.imavis.2020.103981.

[8]. S. S. Ashrafi, S. B. Shokouhi, and A. Ayatollahi, “Action recognition in still images using a multi-attention guided network with weakly supervised saliency detection,” Multimedia Tools and Applications, vol. 80, no. 21–23, pp. 32567–32593, Jul. 2021, doi: 10.1007/s11042-021-11215-1.

[9]. T. Admin, “What object categories / labels are in COCO Dataset?,” Amikelive | Technology Blog, Oct. 29, 2022. https://tech.amikelive.com/node-718/what-object-categories-labels-are-in-coco-dataset/

[10]. B. Kulis and K. Grauman, “Kernelized locality-sensitive hashing for scalable image search,” 2009 IEEE 12th International Conference on Computer Vision, Sep. 2009, doi: 10.1109/iccv.2009.5459466.

[11]. K. Han, J. Guo, C. Zhang, and M. Zhu, “Attribute-Aware Attention Model for Fine-grained Representation Learning,” MM ’18: Proceedings of the 26th ACM International Conference on Multimedia, Oct. 2018, doi: 10.1145/3240508.3240550.

[12]. X. Wang et al., “Pedestrian attribute recognition: A survey,” Pattern Recognition, vol. 121, p. 108220, Jan. 2022, doi: 10.1016/j.patcog.2021.108220.

[13]. H. T. Vu and C.-C. Huang, “Parking space status inference upon a deep CNN and Multi-Task contrastive network with spatial transform,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 4, pp. 1194–1208, Apr. 2019, doi: 10.1109/tcsvt.2018.2826053.

[14]. K. Yang, X. Hu, and R. Stiefelhagen, “Is Context-Aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild,” IEEE Transactions on Image Processing, vol. 30, pp. 1866–1881, Jan. 2021, doi: 10.1109/tip.2020.3048682.

[15]. X. Gao, W. Xu, M. Liao, and G. Chen, “Trust Prediction for Online Social Networks with Integrated Time-Aware Similarity,” ACM Transactions on Knowledge Discovery From Data, vol. 15, no. 6, pp. 1–30, May 2021, doi: 10.1145/3447682.

[16]. M. Ye and Y. Guo, “Zero-Shot Classification with Discriminative Semantic Representation Learning,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, doi: 10.1109/cvpr.2017.542.

[17]. K. Wang, L. Zhang, Y. Tan, J. Zhao, and S. Zhou, “Learning Latent Semantic Attributes for Zero-Shot Object Detection,” 2020 IEEE 32nd International Conference on Tools With Artificial Intelligence (ICTAI), Nov. 2020, doi: 10.1109/ictai50040.2020.00045.

[18]. H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, “Consensus-Aware Visual-Semantic embedding for Image-Text matching,” in Lecture Notes in Computer Science, 2020, pp. 18–34. doi: 10.1007/978-3-030-58586-0_2.

[19]. L. Zhang, J. Rojas, J. Liu, and Y. Guan, "Visual-Semantic Graph attention networks for Human-Object Interaction Detection," 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Dec. 2021, doi: 10.1109/robio54168.2021.9739429.

[20]. C. Gao, Y. Zou, and J.-B. Huang, "ICAN: Instance-Centric Attention Network for Human-Object Interaction Detection.," arXiv (Cornell University), p. 41, Jan. 2018, [Online]. Available: http://arxiv.org/pdf/1808.10437.pdf

[21]. S. Vahora and N. Chauhan, "Deep neural network model for group activity recognition using contextual relationship," Engineering Science and Technology, an International Journal, vol. 22, no. 1, pp. 47–54, Feb. 2019, doi: 10.1016/j.jestch.2018.08.010.

[22]. H. Liu, T. J. Mu, and X. Huang, "Detecting human—object interaction with multi-level pairwise feature network," Computational Visual Media, vol. 7, no. 2, pp. 229–239, Oct. 2020, doi: 10.1007/s41095-020-0188-2.

[23]. H. Wang, Y. Huang, and Q. Zhang, "Human-Object Interaction Detection via Global Context and Pairwise-Level Fusion Features Integration," Neural Networks, Jan. 2022, doi: 10.2139/ssrn.4299944.

[24]. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, doi: 10.1109/cvpr.2017.106.

[25]. S.-W. Kim, H. K. Kook, J. Y. Sun, M. C. Kang, and S. J. Ko, "Parallel feature Pyramid Network for object detection," in Lecture Notes in Computer Science, 2018, pp. 239–256. doi: 10.1007/978-3-030-01228-1_15.

[26]. P. Keserwani, A. Dhankhar, R. Saini, and P. P. Roy, "Quadbox: Quadrilateral bounding box based scene text detection using vector regression," IEEE Access, vol. 9, pp. 36802–36818, Jan. 2021, doi: 10.1109/access.2021.3063030.

[27]. Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A Benchmark for Recognizing Human-Object Interactions in Images," 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 2015, doi: 10.1109/iccv.2015.122.

[28]. D.-T. Le, J. Uijlings, and R. Bernardi, "TUHOI: Trento Universal Human Object Interaction Dataset," Proceedings of the Third Workshop on Vision and Language, Jan. 2014, doi: 10.3115/v1/w14-5403.

[29]. R. Kumar, B. Lahiri, and A. Kr. Ojha, "Aggressive and Offensive language identification in Hindi, Bangla, and English: A Comparative study," SN Computer Science, vol. 2, no. 1, Jan. 2021, doi: 10.1007/s42979-020-00414-6.

[30]. Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," Remote Sensing of Environment, vol. 250, p. 112012, Dec. 2020, doi: 10.1016/j.rse.2020.112012.

[31]. J. Shen, Y. Zhou, Y. Wang, X. Chen, T. Han, and T. Chen, "Evaluating Code Summarization with Improved Correlation with Human Assessment," 2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS), Dec. 2021, doi: 10.1109/qrs54544.2021.00108.

[32]. O. M. Nezami, M. Dras, P. Anderson, and L. Hamey, "Face-Cap: Image captioning using facial expression analysis," in Lecture Notes in Computer Science, 2019, pp. 226–240. doi: 10.1007/978-3-030-10925-7_14.

[33]. V. Frehe, J. Mehmann, and F. Teuteberg, "Understanding and assessing crowd logistics business models – using everyday people for last mile delivery," Journal of Business & Industrial Marketing, vol. 32, no. 1, pp. 75–97, Feb. 2017, doi: 10.1108/jbim-10-2015-0182.

[34]. L. J. Graham, K. De Bruin, C. Lassig, and I. Spandagou, "A scoping review of 20 years of research on differentiation: investigating conceptualisation, characteristics, and methods used," Review of Education, vol. 9, no. 1, pp. 161–198, Nov. 2020, doi: 10.1002/rev3.3238.

[35]. A. M. Borghi and L. Riggio, "Sentence comprehension and simulation of object temporary, canonical and stable affordances," Brain Research, vol. 1253, pp. 117–128, Feb. 2009, doi: 10.1016/j.brainres.2008.11.064.

[36]. D. Roy, "Learning visually grounded words and syntax for a scene description task," Computer Speech & Language, vol. 16, no. 3–4, pp. 353–385, Jul. 2002, doi: 10.1016/s0885-2308(02)00024-4.

[37]. K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," Computer Vision and Image Understanding, vol. 163, pp. 3–20, Oct. 2017, doi: 10.1016/j.cviu.2017.06.005.

[38]. S. Antol et al., "VQA: Visual Question Answering," 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 2015, doi: 10.1109/iccv.2015.279.

[39]. L. Yu, E. Park, A. C. Berg, and T. L. Berg, "Visual Madlibs: Fill in the Blank Description Generation and Question Answering," 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 2015, doi: 10.1109/iccv.2015.283.

[40]. R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," Machine Learning, vol. 8, no. 3–4, pp. 229–256, May 1992, doi: 10.1007/bf00992696.

[41]. S. Ayub, N. Singh, Md. Z. Hussain, M. Ashraf, D. K. Singh, and A. Haldorai, "Hybrid approach to implement multi-robotic navigation system using neural network, fuzzy logic, and bio-inspired optimization methodologies," Computational Intelligence, vol. 39, no. 4, pp. 592–606, Sep. 2022, doi: 10.1111/coin.12547.

[42]. Haldorai, Ravishankar. C. V, Q. S. Mahdi, and G. J. Nehru, "Advanced Communication in Cyber Physical System Infrastructure, Protocols, and Challenges," 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Feb. 2023, doi: 10.1109/icecct56650.2023.10179710.

[43]. J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, Jul. 2003, doi: 10.1145/860435.860459.

[44]. P. Gupta, R. E. Banchs, and P. Rosso, "Continuous space models for CLIR," Information Processing and Management, vol. 53, no. 2, pp. 359–370, Mar. 2017, doi: 10.1016/j.ipm.2016.11.002.

[45]. L. Zheng, C. Zhang, and C. Chen, "MMDF-LDA: An improved Multi-Modal Latent Dirichlet Allocation model for social image annotation," Expert Systems With Applications, vol. 104, pp. 168–184, Aug. 2018, doi: 10.1016/j.eswa.2018.03.014.

[46]. T. Melzer, M. J. Reiter, and H. Bischof, "Appearance models based on kernel canonical correlation analysis," Pattern Recognition, vol. 36, no. 9, pp. 1961–1971, Sep. 2003, doi: 10.1016/s0031-3203(03)00058-x.

[47]. Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving Image-Sentence embeddings using large weakly annotated photo collections," in Lecture Notes in Computer Science, 2014, pp. 529–545. doi: 10.1007/978-3-319-10593-2_35.

[48]. H. Müller, N. Michoux, D. Bandon, and A. Geissbühler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," International Journal of Medical Informatics, vol. 73, no. 1, pp. 1–23, Feb. 2004, doi: 10.1016/j.ijmedinf.2003.11.024.

[49]. A. Pimpalkar and J. R. Raj, "MBiLSTMGloVe: Embedding GloVe knowledge into the corpus using multi-layer BiLSTM deep learning model for social media sentiment analysis," Expert Systems With Applications, vol. 203, p. 117581, Oct. 2022, doi: 10.1016/j.eswa.2022.117581.