EUROPEAN JOURNALS OF EMERGING COMPUTER VISION AND NATURAL LANGUAGE PROCESSING

REAL-TIME AUDITORY GUIDANCE FOR THE VISUALLY IMPAIRED: AN F-RCNN APPROACH IN ASSISTIVE ROBOTICS

Dr. Eryndor V. Mallin
School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

Dr. Taisia L. Khoren
Department of Electrical Engineering and Information Technology, ETH Zurich, Switzerland

V0LUME01 ISSUE01 (2024)

Published Date: 30 December 2024 // Page no.: - 101-108

ABSTRACT

The convergence of Computer Vision (CV) and Natural Language Processing (NLP), two of the most dynamic research areas in machine learning, is charting a transformative course for the field of robotics. This article delves into the intricate integration of these two pivotal domains of artificial intelligence to enhance the capabilities of multimedia robotics applications. We explore how robots, by simultaneously interpreting visual data from their environment and comprehending human language, can achieve unprecedented levels of interaction and operational sophistication. The discussion navigates through the foundational principles of CV and NLP, highlighting the evolution of techniques from classical methods to advanced deep learning models [9]. We examine the methodologies behind fusing visual and linguistic data, focusing on architectures that enable robots to perform complex tasks such as object recognition and manipulation based on verbal commands [14]. A significant focus is placed on a practical application of this synergy: an assistive technology for visually impaired individuals, which utilizes a smartphone paired with a Faster Region Convolutional Neural Network (F-RCNN) based server to identify obstacles and provide real-time auditory guidance. This article presents an in-depth analysis of the applications, benefits, and inherent challenges of this integration, drawing upon a wide array of research. Through a comprehensive review of existing literature, we illustrate the profound impact of this synergy on creating more intelligent, autonomous, and intuitive robotic systems. The findings suggest that the continued advancement in the fusion of CV and NLP will be instrumental in realizing the full potential of social and industrial robots in our society [1].

Keywords: Computer Vision, Natural Language Processing, Multimedia Robotics, Human-Robot Interaction, Deep Learning, Data Fusion, Assistive Technology, F-RCNN, Visually Impaired.

INTRODUCTION

In the landscape of modern science and technology, the principles of integration and interdisciplinarity are paramount for addressing complex, real-world challenges. Nowhere is this more evident than in the field of Artificial Intelligence (AI), where the fusion of distinct sub-disciplines is unlocking previously unattainable capabilities. This article focuses on the synergy between two such areas: Computer Vision (CV) and Natural Language Processing (NLP), and its profound application in the domain of multimedia and robotics. While both CV and NLP are active and powerful research fields in their own right, their integration gives rise to a new interdisciplinary frontier that is attracting significant research attention.

Humans effortlessly use their vision to perceive the world and language to communicate their understanding to others; we can look at a scene and describe it, or read a text and form a mental image. For machines, however, these are monumental tasks that require a deep synthesis of knowledge from both domains. The convenience of interacting with robots through natural gestures and spoken language is a primary driver for this research, as it represents the most intuitive form of human-robot interaction. This potential can only be realized if robots are capable of understanding these complex, multimodal inputs. The real-world need for this integration is already apparent in our daily lives, from subtitled videos to images tagged on social media, where visual and textual data streams are intrinsically linked to convey a complete message.

1.1 The Domain of Computer Vision

Computer Vision (CV) is the scientific endeavor focused on enabling machines to perceive, process, and understand visual data from the world in a manner similar to humans. The application of vision capabilities is rapidly expanding across industries, powering systems for surveillance, medical imaging, autonomous vehicles, object recognition, and activity monitoring1212. The core tasks within computer vision can be broadly understood through the "three R's":

- Recognition: This involves the identification and labeling of objects within an image. Examples range from face and handwriting recognition to the broader task of general object recognition. It is the foundational ability that allows a robot to know what it is looking at.

- Reconstruction: This task aims to estimate a three-dimensional (3D) model of a scene from two-dimensional (2D) images. It leverages information from various cues like multiple camera views, shading, texture, or direct depth sensors to provide the robot with a spatial understanding of its environment.

- Reorganization: Often referred to as bottom-up vision, this involves structuring an image by grouping raw pixels into meaningful segments. This includes fundamental processes like edge detection, corner detection, and semantic segmentation, which is the act of partitioning an image into semantically meaningful parts.

1.2 The Domain of Natural Language Processing

Natural Language Processing (NLP) is the branch of AI that empowers computers with the ability to understand, interpret, and generate human language, whether in spoken or written form. The ultimate goal is to enable machines to produce and comprehend language with the same fluency and nuance as humans [5, 53]. This involves a hierarchy of complex tasks:

- Morphological and Syntactic Analysis: This is the foundational level of understanding the structure of language. It involves breaking words down into their basic components (morphology) and analyzing the grammatical structure of sentences (syntax)21.

- Semantic Analysis: This phase focuses on extracting the meaning from words and sentences, moving beyond literal structure to understand the concepts being conveyed.

- Discourse and Pragmatic Analysis: These are higher-level tasks. Discourse analysis involves understanding the context and flow of language beyond a single sentence 23, while pragmatic analysis deals with the intended meaning, which can often differ from the literal meaning based on social context and real-world knowledge.

Complex NLP applications include machine translation, automated summarization, and interactive dialogue systems [10, 71].

1.3 The Convergence of Vision and Language

The true power emerges when these two fields are intertwined. The connection between CV and NLP is intrinsically semantic [7, 68]. The outputs of computer vision tasks naturally map to linguistic components: recognized objects can be described by

nouns, detected activities and actions by verbs, and the properties or features of those objects by adjectives. This semantic linkage is the bridge that allows a machine to not just "see" an object, but to "describe" it as a "large red ball" [8]. Conversely, NLP can be used to guide computer vision. Describing an image in words can be seen as a form of machine translation, where the task is to translate low-level pixel data into a high-level, human-readable description [9, 72].

The scope for this integrated approach is vast, primarily because so much of modern data is inherently multimedia, containing both visual and textual information that complement each other to tell a single, clearer story2626. This interdisciplinary approach is proving extremely useful for tasks like object and text recognition in multimedia and robotics applications [10, 79].

1.4 Core Challenges and Motivations

Despite the immense potential, the integration of CV and NLP is fraught with significant challenges that motivate ongoing research [2, 12].

For Computer Vision, one of the most formidable obstacles is the creation of large-scale, high-quality datasets. Traditional machine learning and deep learning approaches require vast amounts of labeled, annotated, and segmented data. This process is often performed manually by human experts and is incredibly time-consuming and expensive [11]. The challenge is compounded by the inherent variability of the real world. An object can appear drastically different depending on factors like illumination, viewing angle, scale, orientation, and occlusion. Manually annotating every possible variation of every object is a phenomenal and inefficient task.

For Natural Language Processing, while significant progress has been made in syntactic analysis, major hurdles remain in the semantic and pragmatic realms. Applications that can robustly handle word sense disambiguation (determining the correct meaning of a word in context), deep semantic analysis, and pragmatic understanding are still in high demand. The complexity is magnified by the sheer number of languages and dialects across the globe, each with its own unique grammar and cultural context.

These challenges motivate the development of integrated models where vision and language can jointly supervise and balance each other, bridging the gaps that exist in each individual domain [13, 92]. The current work explores these integrated methodologies, with a particular focus on their application in a system designed to guide the movements of visually impaired individuals, offering a tangible example of how this technology can provide accurate, automated assistance without the need for human intervention3434.

2. Methodologies of Integration

The successful integration of computer vision and natural language processing in robotics is not a matter of simply connecting two separate modules. It requires the development of sophisticated methodologies for representation, data fusion, and grounding [13]. These methods aim to create a cohesive system where the robot can reason jointly about what it sees and what it hears or reads. This section explores the foundational models, fusion techniques, and a specific system design that exemplifies this integration.

2.1 Foundational Models and Representations

The bedrock of modern CV and NLP integration is deep learning, which provides powerful tools for learning rich data representations [9].

- Visual Representation: For the visual domain, Convolutional Neural Networks (CNNs) are the dominant architecture [39]. Models like VGG, ResNet, and Inception are trained on massive image datasets (e.g., ImageNet) to learn a hierarchical set of features. The output is typically a high-dimensional feature vector that encodes the visual content of an image or video frame. These networks form the "eyes" of the robotic system. An important development in this area is the use of a deep, fully convolutional network as a

Region Proposal Network (RPN), which can efficiently propose candidate object regions within an image for further analysis.

- Linguistic Representation: For the linguistic domain, the challenge is to convert symbolic text into continuous vector representations. Early methods relied on one-hot encodings, but these were sparse and lacked semantic meaning. The breakthrough came with word embeddings, such as WordVec and GloVe, which learn dense vectors for words where semantically similar words are closer in the vector space [25, 40]. To process sequences of words, Recurrent Neural Networks (RNNs) and their more advanced variants, Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, became the standard [33]. These models process input sequentially, maintaining a hidden state that captures information from previous inputs. More recently, Transformer models, which use attention mechanisms instead of recurrence, have shown state-of-the-art performance on a wide range of NLP tasks [9, 10].

- Shared Multimodal Space: The ultimate goal is to project these distinct visual and linguistic representations into a common, shared space where they can be meaningfully compared and combined [13]. This is often achieved through a structured objective function that aligns the two modalities during training [24]. For example, a model can be trained to ensure that the vector representation of an image of a dog is close to the vector representation of the sentence "a photo of a dog" [24]. This deep visual-semantic alignment is a cornerstone of many modern image description and retrieval models. Some approaches refine this by moving beyond raw image features to use visual attributes (e.g., "is red," "has fur," "is metallic"), as attributes are not confined to specific object categories and can differentiate concepts more clearly [27, 174].

2.2 Data Fusion and Grounding Techniques

Once representations are learned, they must be fused. This can be approached from different directions:

- Top-Down vs. Bottom-Up Approaches: A top-down approach uses a language model to guide the visual analysis, essentially generating a description and then trying to ground it in the image. A bottom-up approach starts with detecting salient objects and features in the image (e.g., using keywords) and then composes them into a sentence. Many successful systems use a hybrid model that combines the strengths of both, perhaps using a bottom-up method for initial concept detection and a top-down method for generating fluent language [22, 148, 149].

- Fusion Models: Simple concatenation of visual and language vectors is a common baseline. More sophisticated techniques aim for a richer interaction between modalities.

Canonical Correlation Analysis (CCA) is a statistical method used to maximize the correlation between the two sets of vectors. An extension, the

The 3-view CCA model incorporates a third view representing high-level semantic information (e.g., from a knowledge base), which helps to better separate object classes and significantly improves image retrieval accuracy compared to the standard 2-view approach [23, 154, 156].

- Grounding Natural Language: "Grounding" refers to the critical process of connecting abstract language symbols to the physical objects, attributes, and spatial relationships in the robot's perceived environment [14]. This is what allows a robot to understand that the phrase "the blue cup on the table" refers to a specific entity in its field of view. Probabilistic graphical models, such as

Generalized Grounding Graphs (G), have been proposed to achieve this. These models represent the structure of a natural language command, mapping linguistic elements (like spatial description clauses) to features of grounding such as objects, places, or paths. By training on a corpus of commands paired with their correct groundings, the model can automatically learn the meaning of words in a physically grounded context [14, 101].

2.3 A Case Study in Assistive Technology: System Design for the Visually Impaired

To make these concepts concrete, we can examine the design of an integrated system proposed to assist visually impaired individuals, as detailed in the source material. This system serves as an excellent case study for applying the aforementioned methodologies to a real-world problem.

The overall architecture is user-friendly and low-cost, relying on a standard smartphone4141. The blind user holds the smartphone, which continuously captures images of their surroundings. These images are processed to recognize objects and potential obstacles, and the results are communicated back to the user as spoken guidance. The system is composed of two primary modules: a computer vision module for object recognition and a natural language processing module for generating the spoken output.

- Computer Vision Module (Faster R-CNN): For the demanding task of real-time object recognition, the system utilizes the Faster R-CNN (F-RCNN) algorithm. F-RCNN is a two-stage detector known for its accuracy and relative speed.

1. First, the input image is passed through a series of convolutional and max-pooling layers to produce a high-level feature map.

2. Next, a Region Proposal Network (RPN), which is a deep fully convolutional network, slides over this feature map and proposes a set of rectangular object proposals, each with an "objectness" score4646.

3. These proposals, which are Regions of Interest (ROIs), are then used to pool features from the feature map, creating a fixed-length feature vector for each proposal.

4. Finally, this vector is fed into a classifier (the Fast R-CNN detector module) which determines the specific class of the object (e.g., "chair," "table," "person") and refines the bounding box. The output is a list of detected objects and their locations in the image.

- Natural Language Processing Module: The object recognition output (e.g., a list of labels like ['chair', 'door']) is then passed to the NLP module to be converted into a useful, human-understandable message. This involves several stages:

1. Text Pre-processing: Before any language modeling, the text data is rigorously cleaned to improve efficiency and accuracy. This involves removing any extraneous characters like HTML tags, which provide no semantic value but consume processing time; stripping away punctuation; removing common but uninformative "stop words" like 'is', 'a', 'the'; and performing stemming, which reduces words to their root form (e.g., 'moving' and 'moved' become 'move') to avoid creating redundant vectors. Finally, all text is converted to lowercase to ensure words like 'Chair' and 'chair' are treated as identical.

2. Vectorization (Bag-of-Words): Since NLP algorithms work with numbers, the cleaned text must be converted into a numerical format. The system uses the Bag-of-Words (BOG) model for this task. BOG works by creating a vocabulary of all unique words in the text and then representing each sentence as a vector where each element is the count of a particular word's occurrence.

3. Speech Synthesis: The final step is to convert the processed natural language text (e.g., "There is a chair in front of you") into speech. This text is sent from the server back to the user's smartphone, which uses its built-in text-to-speech capabilities to provide auditory guidance to the user, helping them navigate their environment safely.

This complete pipeline, from image capture to spoken guidance, illustrates a powerful and practical application of integrating CV and NLP to solve a meaningful human problem [424].

3. Applications and Performance

The theoretical integration of vision and language translates into a diverse array of practical applications that are expanding the capabilities of robotic systems. These applications range from generating rich, human-like descriptions of the environment to enabling sophisticated knowledge transfer between humans and robots. Furthermore, experimental analysis of the core deep learning models provides insight into their relative strengths and weaknesses across different tasks.

3.1 Key Application Areas

- Image and Video Captioning: One of the most prominent applications is the automatic generation of textual descriptions for visual media [21]. This goes far beyond simple keyword tagging. Modern systems can generate full, grammatically correct sentences that describe the content and context of an image or video [15]. This is achieved through various methods, such as predicting the most probable nouns, verbs, and prepositions that could form a sentence for a given image, or by identifying the most likely subject-verb-object (SVO) triplet to describe the main action in a video [17, 116]. For longer videos, techniques like using latent topics and sparse object stitching can generate concise summaries from thousands of frames [22]. More advanced approaches utilize scene description graphs, which represent the objects, attributes, and relationships in a scene, and then leverage an automatically constructed knowledge base to reason about the scene and generate rich, detailed captions [20, 131, 132]. A comprehensive survey of these models highlights that they can often produce descriptions of a quality comparable to those written by humans [21, 143].

- Robotic Command and Control: A major goal of HRI is to allow users to command robots using natural, everyday language [14]. Integrated systems enable a robot to parse a command like "bring me the green bottle from the kitchen table," identify the objects ("bottle"), their attributes ("green"), and their spatial location ("on the kitchen table"), and then execute the appropriate navigation and manipulation actions [14]. The robot often learns the meaning of words through training on large corpora where commands are paired with their corresponding correct groundings in the world [14, 101].

- Human-Robot Knowledge Transfer: Beyond simple commands, this integration facilitates deeper knowledge transfer [26]. A framework has been demonstrated where a robot learns complex tasks in real-time by observing human demonstrations. The acquired knowledge is represented in a hierarchical Spatial, Temporal, and Causal And-Or Graph (STC-AOG), which functions as a form of stochastic grammar. This allows the robot to build a knowledge base of skills that can be retrieved and adapted, essentially acting as a repository where humans can deposit skills for later use by both robots and other humans [26, 171].

● Enhanced Video Activity Recognition: The combination of vision and language can significantly improve a robot's ability to recognize activities in video [16]. One method uses text mining on large web corpora to learn the correlation between certain verbs (actions) and objects. This linguistic knowledge is then combined with the output of visual object detectors and activity classifiers to enhance the overall accuracy of activity recognition in videos, even without explicit labeling of the training data [16, 114].

3.2 Performance Analysis of Core Models

To understand the effectiveness of the underlying technologies, it's crucial to analyze the performance of the core deep learning architectures. The provided source material details a comparative study of CNNs, GRUs, and LSTMs across a suite of NLP tasks, which can be grouped into broad categories. The experiments were conducted by training each model from scratch with individually tuned hyperparameters to ensure a fair comparison6363.

● Text Classification (e.g., Sentiment Analysis): In these tasks, the goal is to assign a category to a piece of text [33]. Surprisingly, while CNNs are often considered strong at extracting key local features (like specific sentiment-bearing words), RNN-based models (GRU and LSTM) were found to outperform them, particularly on sentiment classification. This suggests that for tasks like sentiment analysis, understanding the full compositional structure of the sentence, which RNNs are adept at, is often more important than just identifying keywords.

● Semantic Matching (e.g., Answer Selection, Textual Entailment): These tasks require a model to understand the semantic relationship between two pieces of text [35]. Here, the results were mixed. CNNs performed better on some tasks like Answer Selection (AS) and Question Relation Match (QRM), while RNNs were superior on Textual Entailment (TE)66. This indicates that the optimal architecture depends heavily on the specific nature of the semantic comparison being made.

● Sequence Ordering (e.g., Path Query Answering): For tasks that rely heavily on understanding long-range order and dependencies, such as Path Query Answering (PQA) on a knowledge base, both GRU and LSTM significantly outperformed CNNs. This aligns with theoretical expectations, as the recurrent nature of RNNs is explicitly designed to model sequential information.

● Context Dependency (e.g., Part-of-Speech Tagging): In this category, CNNs were found to be superior to standard unidirectional RNNs. However, they fell short of bi-directional RNNs. This makes sense, as PoS tagging for a given word often depends on the context from both the left and right, which a bi-directional model can capture, whereas a standard CNN has a fixed-size receptive field.

A key takeaway from this analysis is that there is no single "best" model. RNNs (GRU and LSTM) are highly resilient and perform well across a wide range of tasks, especially when the input's structural and long-range dependency information is important. A detailed error analysis on sentiment classification revealed that GRU handles sentences with inverted semantic clauses (e.g., "this edition is not classic... but its delights are still ample") better than CNNs, especially as sentence length increases. However, for sentences where a few local keywords determine the outcome, the ability of RNNs to model the entire phrase can sometimes be a disadvantage, as important local details might be overlooked. The performance of all models was also shown to be highly sensitive to hyperparameters like hidden size and batch size, highlighting the critical importance of proper tuning for achieving optimal results.

4. Discussion and Future Directions

The rapid advancements in integrating computer vision and natural language processing have brought us closer to the goal of truly intelligent and collaborative robots. The field's progress can be characterized by three key themes: accuracy, scalability, and innovation.

Accuracy has been dramatically improved by the success of deep learning models like CNNs and RNNs.

Scalability has been enabled by breakthroughs in high-performance computing and hardware acceleration, allowing these complex models to be trained on massive datasets.

Innovation is evident in the wealth of novel applications that have emerged, from video captioning to conversational systems.

However, despite this significant progress, a deeper discussion reveals critical limitations and points to important future research directions. The overarching challenge can be framed as bridging the vast semantic gap—the chasm between low-level sensory data and high-level, human-like conceptual understanding.

4.1 The Remaining Semantic Gap

Current systems, while impressive, often operate on a relatively shallow level of understanding. A deep learning system trained on millions of images can accurately classify a vast number of object categories, far surpassing earlier methods. Yet, this is still a far cry from true visual acuity or comprehension. When a human sees a bird, the sensory input immediately activates a rich network of background knowledge. We know that the bird is likely alive, that it probably has the ability to fly (while implicitly recognizing that a broken wing could prevent this), and we can place this observation within a broader context of past experiences and world knowledge.

Today's frameworks are largely incapable of this kind of effective reasoning. An object detector may identify a "person" and a "knife" in a kitchen scene, but it lacks the commonsense knowledge to distinguish between someone chopping vegetables and a threatening situation. This fine-grained, context-aware understanding requires a new class of model architectures that can perform scaled structural prediction and integrate vast stores of background knowledge8282. The development of distributional semantics, using everything from classic bag-of-words methods to modern wordvec embeddings, is a step in this direction, but it is not enough [34, 40, 350].

4.2 Architectural and Methodological Future Directions

Addressing these fundamental limitations will require innovation on multiple fronts.

- Handling Complex Data Structures: Standard RNNs and LSTMs excel at processing sequential data, but much of the world's information is not purely sequential. Language has a hierarchical, tree-like structure, and scenes have complex spatial and causal relationships. When the input data structure is not known ahead of time, a standard LSTM can fail. Future research must focus on architectures that can handle more complex data structures. Extensions like

Tree-LSTM, which is designed to handle tree-structured data, or even more flexible graph neural networks, offer a promising path forward for better modeling both linguistic syntax and visual scene graphs [40].

- Tensor-Based Multimodal Semantics: A powerful future direction for modeling the rich, bidirectional interactions between vision and language is tensor decomposition. A tensor can represent a wide range of multimodal information, such as word co-occurrence data for topic modeling or graph adjacency data for network analysis. Just as Singular Value Decomposition (SVD) can decompose a matrix into a set of orthogonal bases, tensor decomposition methods can deconstruct a multimodal data tensor into its most relevant underlying statistical components. This could provide a principled and powerful mathematical foundation for the next generation of multimodal distributional semantics.

- Attention and Memory: The success of Transformer models has shown the power of attention mechanisms. Further research into attention, particularly for systems that must integrate long-term memory with immediate sensory input, is crucial. A robot should be able to pay attention to the most relevant parts of a visual scene while simultaneously recalling past interactions or instructions related to that scene.

In essence, the next great leap will likely come from moving beyond pattern recognition towards models that incorporate structure, memory, and reasoning.

5. CONCLUSION

The integration of computer vision and natural language processing is undeniably one of the most exciting and consequential frontiers in robotics and artificial intelligence. This article has explored the multifaceted relationship between these two domains, from the foundational deep learning models that power them to the sophisticated methodologies used to weave them together. We have discussed their application to a wide range of tasks in multimedia and robotics, from image captioning to intuitive human-robot interaction [3, 13, 28]. Through a detailed case study of a smartphone-based guidance system for the visually impaired, we have seen how this synergy can be harnessed to create simple, low-cost, and user-friendly solutions to profound real-world problems, significantly improving the mobility and independence of users8989.

Our review has shown that while remarkable progress has been made, the journey towards true machine intelligence is far from over. Comparative analyses demonstrate that while RNNs are robust performers across many NLP tasks, no single architecture is a panacea; the choice of model remains highly task-dependent. More importantly, a significant semantic gap persists between the pattern recognition capabilities of current systems and the deep, context-aware understanding that characterizes human cognition9191.

The future of the field lies in bridging this gap. This will require a concerted push towards new model architectures that can handle complex, non-sequential data structures, frameworks for integrating vast stores of commonsense and background knowledge, and principled mathematical approaches for modeling rich multimodal data [20, 404, 410]. By continuing to push these boundaries, the research community will move closer to creating the next generation of robots—machines that not only see our world and hear our words, but truly understand them.

REFERENCES

1. G. Yin, Intelligent framework for social robots based on artificial intelligence-driven mobile edge computing, Computers & Electrical Engineering, 96, Part B, (2021).
2. Fisher, M., Cardoso, R. C., Collins, E. C., Dadswell, C., Dennis, L. A., Dixon, C., ... & Webster, M., An overview of verification and validation challenges for inspection robots, Robotics, 10, 67 (2021).
3. A. Jamshed and M. M. Fraz, NLP Meets Vision for Visual Interpretation - A Retrospective Insight and Future directions, 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT), 1-8 (2021).

4. W. Fang, P. E.D. Love, H. Luo, L. Ding, Computer vision for behaviour-based safety in construction: A review and future directions, Advanced Engineering Informatics, 43, (2020).

5. H. Sharma, Improving Natural Language Processing tasks by Using Machine Learning Techniques, 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 1-5 (2021).

6. M. Jitendra, P. Arbeláez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani, The three R's of computer vision: Recognition, reconstruction and reorganization, Pattern Recognition Letters, 72, 4-14 (2016).

7. P. Gärdenfors, The Geometry of Meaning: Semantics Based on Conceptual Spaces, MIT Press, (2014).

8. E. Dockrell, D. Messer, R. George, and A. Ralli, Beyond naming patterns in children with WFDs—Definitions for nouns and verbs, Journal of Neurolinguistics, 16, 191-211 (2003).

9. A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, Natural language processing advancements by deep learning: A survey, arXiv preprint arXiv:2003.01200 (2020).

10. W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, and C. Santos-Libarino, Emotion detection for social robots based on nlp transformers and an emotion ontology, Sensors, 21, 1322 (2021).

11. S., Zhenfeng, W. Wu, Z. Wang, W. Du, and C. Li, Seaships: A large-scale precisely annotated dataset for ship detection, IEEE transactions on multimedia, 20, 2593-2604 (2018).

12. https://monkeylearn.com/blog/natural-language-processing-challenges/ , last vist 1/2/2022.

13. C. Zhang, Z. Yang, X. He and L. Deng, Multimodal Intelligence: Representation Learning, Information Fusion, and Applications, in IEEE Journal of Selected Topics in Signal Processing, 14, 478-493 (2020).

14. S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy, Understanding natural language commands for robotic navigation and mobile manipulation. In Proceedings of the AAAI Conference on Artificial Intelligence, 25, 1507-1514 (2011).

15. Y. Yezhou, C. Teo, H. Daumé III, and Y. Aloimonos, Corpus-guided sentence generation of natural images, In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 444-454 (2011).

16. T. S. Motwani, R. J. Mooney, Improving Video Activity Recognition using Object Recognition and Text Mining, In Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012), 600-605 (2012).

17. N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko and S. Guadarrama, Generating Natural-Language Video Descriptions Using Text-Mined Knowledge, In Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI-2013), 541-547 (2013).

18. J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R Mooney, Integrating language and vision to generate natural language descriptions of videos in the wild, Proceedings of the 25th International Conference on Computational Linguistics (COLING), (2014).

19. Y. Yezhou, C. L. Teo, C. Fermüller, and Y. Aloimonos, Robots with language: Multi-label visual recognition using NLP, In IEEE International Conference on Robotics and Automation, 4256-4262 (2013).

20. S. Aditya, Y. Yang, C. Baral, C. Fermuller, and Y. Aloimonos, From images to sentences through scene description graphs using commonsense reasoning and knowledge, arXiv preprint arXiv, (2015).

21. R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. I. Cinbis, F. Keller, A. Muscat, and B. Plank, Automatic description generation from images: A survey of models, datasets, and evaluation measures, Journal of Artificial Intelligence Research, 55, 409-442 (2016).

22. P. Das, C. Xu, R. Doell, and J. Corso, A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2634-264 (2013).

23. Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, International journal of computer vision, 106, 210-233 (2014).

24. A. Karpathy and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3128-3137 (2015).

25. R. Schwartz, R. Reichart and A. Rappoport, Symmetric pattern based word embeddings for improved word similarity prediction, In CoNLL, 2015, 258-267 (2015).

26. N. Shukla, C. Xiong, and S. C. Zhu, A unified framework for human-robot knowledge transfer, In Proceedings of the 2015 AAAI Fall Symposium Series, (2015).

27. Carina Silberer, Vittorio Ferrari, and Mirella Lapat, Models of semantic representation with visual attributes, In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 572-582 (2013).

28. R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2, 207-218 (2014).

29. M. Tapaswi, M. B¨auml, and R. Stiefelhagen, Bookmovie: Aligning video scenes with book chapters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1827–1835 (2015).

30. I. Abdalla Mohamed, A. Ben Aissa, L. F. Hussein, Ahmed I. Taloba, and T. kallel, A new model for epidemic prediction: COVID-19 in kingdom saudi arabia case study", Materials Today: Proceedings, (2021).

31. Ahmed. I. Taloba and S. S. I. Ismail, An Intelligent Hybrid Technique of Decision Tree and Genetic Algorithm for E-Mail Spam Detection, Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 99-104 (2019).

32. Ahmed I. Taloba, M. R. Riad and T. H. A. Soliman, Developing an efficient spectral clustering algorithm on large scale graphs in spark, Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), 292-298 (2017).

33. D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1422–1432.

34. Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in International conference on machine learning, 2014, pp. 1188–1196.

35. S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," ArXiv Prepr. ArXiv05326, 2015.

36. Y. Yang, W. Yih, and C. Meek, "Wikiqa: A challenge dataset for open-domain question answering," in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2013–2018.

37. W. Yih, M. Richardson, C. Meek, M.-W. Chang, and J. Suh, "The value of semantic parse labeling for knowledge base question answering," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 201–206.

38. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," ArXiv Prepr. ArXiv01847, 2016.

39. N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," ArXiv Prepr. ArXiv2188, 2014.

40. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Adv. Neural Inf. Process. Syst., vol. 26, 2013.