Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/110309

Integration of Computer Vision and Natural Language Processing in Multimedia Robotics Application

Amir El-Komy [1,*], *Osama R. Shahin* [2,3], *Rasha M. Abd El-Aziz* [2,4] *and Ahmed I. Taloba* [2,5,*]

[1]English Language Department, College of Science and Arts in Qurayyat, Jouf University, Saudi Arabia
[2]Computer Science Department, College of Science and Arts in Qurayyat, Jouf University, Saudi Arabia
[3]Physics and Mathematics Department, Faculty of Engineering, Helwan University, Egypt
[4]Computer Science Department, Faculty of Computers and Information, Assiut University, Egypt
[5]Information System Department, Faculty of Computers and Information, Assiut University, Egypt

Received: 21 Feb. 2022, Revised: 22 Mar. 2022, Accepted: 24 Mar. 2022.
Publishedonline:1 May 2022.

Abstract: Computer vision and natural language processing (NLP) are two active machine learning research areas. However, the integration of these two areas gives rise to a new interdisciplinary field, which is currently attracting more attention of researchers. Research has been carried out to extract the text associated with an image or a video that can assist in making computer vision effective. Moreover, researchers focus on utilizing NLP to extract the meaning of words through the use of computer vision. This concept is widely used in robotics. Although robots should observe the surroundings from different ways of interactions, natural gestures and spoken languages are the most convenient way for humans to interact with the robots. This would be possible only if the robots can understand such types of interactions. In the present paper, the proposed integrated application is utilized for guiding vision-impaired people. As vision is the most essential in the life of a human being, an alternative source that helps in guiding the blind in their movements is highly important. For this purpose, the current paper uses a smartphone with the capabilities of vision, language, and intelligence which has been attached to the blind person to capture the images of their surroundings, and it is associated with a Faster Region Convolutional Neural Network (F-RCNN) based central server to detect the objects in the image to inform the person about them and avoid obstacles in their way. These results are passed to the smartphone which produces a speech output for the guidance of the blinds.

Keywords: Computer Vision, Natural Language Processing, Robotics, Visually Impaired, Smartphone, Faster-RCNN algorithm.

1 Introduction

In modern science, integration and interdisciplinarity are the key to discovering solutions for several challenging tasks. The current paper focuses on two active research areas, computer vision and natural language processing, for their applications in multimedia and robotics. Although they are separately the most active areas in AI, they can be highly beneficial when integrated [1].

Human beings use their capability of vision to observe the world and their language ability to communicate with others, such as seeing an image and describing it in their language and similarly seeing a text and representing it as an image. These are common activities for humans, but they are challenging tasks for robots [2]. A lot of knowledge about computer vision and NLP is required to achieve such tasks. Both these fields are under Artificial Intelligence (AI) and are recent active research areas separately. But research in any of the two areas is not much advantageous. There are several exciting applications in both fields separately, but there are no such applications combining both fields, even though there are several real-world problems that need expertise in the combined application of both vision- and text-related data, for instance, videos with subtitles, images tagged on social media, etc. [3].

1.1 Computer Vision

Computer vision makes it possible for machines to perceive the world as humans do. To achieve this, it is crucial to process and understand the visual data. The incorporation of vision capabilities has been increasing in current industrial applications, including image rendering, surveillance systems, object recognition, and activity

*Corresponding author e-mail: aitaloba@ju.edu.sa

recognition. Several activities of computer vision exist based on the application [4]. Figure 1 shows the activities of computer vision system.

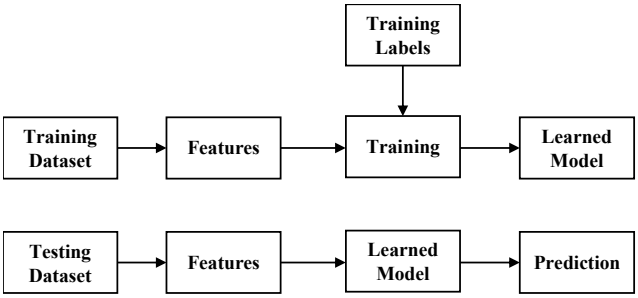

Fig. 1: Activities involved in computer vision system.

1.2 Natural Language Processing (NLP)

Natural language processing means extracting the text from natural language input from a spoken or written communication. The use of NLP is to make the computers generate a sentence from the natural language like humans do [5]. Figure 2 shows the tasks included in the NLP application.

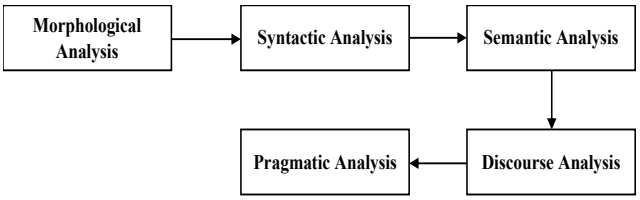

Fig. 2: Tasks involved in natural language processing.

1.3 Computer Vision and Its Relation to NLP

The tasks involved in computer vision are reconstruction, recognition, and reorganization. Reconstruction means estimating a 3D scene that creates an image with the help of the information obtained from multiple views, texture, shading, or direct depth sensors. Recognition indicates labeling of the objects in an image, for example, face recognition, handwriting recognition, and object recognition. Reorganization refers to the bottom-up vision, when an image structure is represented by segmentation of raw pixels, for instance, edge detection, corner detection, contour detection, and semantic segmentation [6].

The way NLP is connected to computer vision indicates that the above-given actions all produce text output. For instance, objects are characterized by nouns, activities are represented by verbs, and features are indicated by adjectives. This way, computer vision and language processing are associated via semantic representation [7-8].

1.4 Natural Language Processing and Its Relation to Computer Vision

NLP is a completely different approach in comparison with computer vision which includes syntax, words, sentences, pragmatics, phrases, and discourses as natural communication. Tasks including summarization, information extraction, dialog interface, and machine translation are considered complex. A task similar to machine translation, which is describing an image in words, involves the translation of low-level pixels of an image to a high-level description which needed more improvement [9].

1.5 Scope of Integration of Computer Vision and NLP

The top-down approach will not be smoother in integrating vision and language. Hence, researchers utilize bottom-up approach, since some pioneers found satisfactory results for certain problems related to it, by attempting multiple solutions. The scope of this interdisciplinary approach is high because most of the files used nowadays are multimedia files containing some natural language texts along with some related images or videos for references.

For instance, a news article comprises some text content written by journalists and an image for the reference to where that event takes place or a video that reveals the truth behind the particular news. From this, it is understood that visual and language data are the two pieces of information that can provide one news, to convey the content clearly with great understanding and without any mistakes to the users. This interdisciplinary approach is extremely useful in recognizing objects and texts and returning their meanings in multimedia and robotics as well [10].

1.6 Challenges and Issues in Computer Vision and NLP

The biggest challenge faced by computer vision is the creation of a dataset. The traditional machine learning approach requires annotation, labeling, segmentation, and so forth, which are manually done by human experts and are considered a huge task when creating a complete dataset with a larger number of real-world objects for different applications. Based on the conditions in the surroundings of an object such as high or low illumination, different viewpoints, or changes in the orientation, scale, or position, such object will look completely different from its original appearance. Hence, manual annotation, labeling, and segmentation by human experts for each object with a massive number of variations are phenomenal tasks and are considered to be not that efficient since they are done by humans [11].

On the other hand, natural language processing is considerably effective when syntax and morphological

analysis are in concern [12]. But when other phases of languages are considered, several glitches arise. There is still a high demand for applications that can handle the problems related to the word sense disambiguation, semantic analysis, and pragmatic analysis. These considerations are highly challenging since there are several languages across nations.

Therefore, integrating vision and natural language application should be developed in such a way that it should bridge the above gap through the use of the combined form of text and image or video and can jointly balance and oversee each other [13].

In this paper, computer vision and natural language processing are integrated in such a way that they can guide the movements of blind people. The usual guidance for blind movements is by the walking stick, then by electronic orientation aid, electronic travel aid, and position locator device. But the vision and language-based movement guide can provide perfect assistance to the blinds accurately and automatically with no human intervention. The present paper is structured as follows: related works are presented in Section 2, followed by the proposed methodology and design in Section 3, and then discussion in Section 4. Finally, the paper concludes in Section 5.

2 Related Works

A model named Generalized Grounding Graphs (G3) is introduced in [14], which is a probabilistic graphical model, represented based on the structure of commands given in natural language. The grounding graph structure is made by a semantic structure called Spatial Description Clauses representing the linguist basics of a command, mapped to the feature of grounding like place, object, event, or path. In a grounding graph, the structure of Spatial Description Clauses and random variables, edges, and nodes depend on the words in the text. Training of this model is performed on a corpus of natural language commands paired with groundings enabling automatic learning of the meaning of the word in the corpus. This model is evaluated on a robotic forklift receiving commands in natural language.

A strategy for sentence generation has been proposed in [15]. In this paper, the sentences are generated by the prediction of the most probable nouns, verbs, prepositions, and scenes that can be used to create a sentence. In an image, with the help of trained detectors, the noisy estimates of objects and scenes are detected and are given as the inputs. It is unreliable to predict the actions directly from a still image. Hence, a language model trained from the English Giga word corpus has been utilized for acquiring their estimates along with the possible nouns, prepositions, and scenes. A Hidden Markov Model uses these estimates as its parameters to generate a sentence.

A method to enhance video activity recognition through the use of object recognition and text mining has been described in [16]. In this method, a combined form of object recognition, activity classification, and text mining has been utilized for learning activity recognizers without obvious labeling of the training videos. The verbs are first clustered to determine the class of the activities automatically and to create a labeled training dataset, which is used for training the activity classifiers based on spatiotemporal features. The correlation of verbs and objects is learned by text mining approach. These data are then used with the outcome of the object recognizer and trained activity classifier to enhance the video activity recognizer.

A data-driven approach for generating natural language descriptions for videos is given in [17]. In this method, the outcomes of object detector and activity detector are used to predict the most likely subject-verb-object triplet to describe a video. The dataset utilized in this approach is the English portion of YouTube containing short videos with various natural language descriptions with 1596 training and 185 testing videos. A discriminatively trained multi-scale, deformable parts model is used as the object detector and motion descriptor for activity detection.

A strategy that can generate the description of a wild-life video has been demonstrated in [18] through the use of a probabilistic factor graph model which utilizes a combination of vision and linguistic information. This model combines the detection confidences on entities, activities, and scenes in a video and the knowledge obtained from text corpora for estimating the most probable Subject (S), Verb (V), Object (O), and Place (P). Furthermore, this method detects the location in which the video has been taken. The dataset used in this method consists of 1297 training and 670 testing videos. The dataset has 45 entities for subjects like person, animal, cat, baby, and chef, 241 entities for objects like tv, person, shrimp, flute, and motorbike, and 218 entities for verbs like walk, ride, play, cut, and climb. The types of sentences generated are SVO, SVP, and SVOP ranked using the BerkeleyLM language model. The output sentence length is finally normalized through the use of the highest average 5-gram probability.

NLP-based multi-label visual recognition for robotics application is given in [19]. This method utilizes two approaches. The first one is creating a database with words related to common daily activities from multiple languages based on different semantic concepts to predict the labels available in a specific context. The second one is the use of statistical language tools that facilitate the correlation of different labels. A large corpus dataset is utilized for learning the linguistic features to provide correlation and integration of the linguistic model with the system. This system is evaluated over 3 multi-label tasks for the recognition of daily activities and results in a considerable improvement in the accuracy with the help of correlation data.

Language description of images can be predicted via extracting the scene description graphs from the scenes in the image, through the use of the automatically created knowledge base by NLP of image annotation. The scene description graph is obtained by vision as well as reasoning that generates the captions for an image. The knowledge

base provides answers for the role of an object in an incident, how the event connects the two objects, and all probable perceptions that involve the objects and the incidents. At last, the scene inference can be obtained by combining the knowledge base with the scene description graph to describe the scene information efficiently. The advantage of using a scene description graph is that it can be processed by AI easily, has rich content, and is not bound to any specific templates for converting the labels to sentences [20].

A survey on the automatic generation of image description is provided in [21]. In this survey, a detailed description of the datasets, models, and evaluation techniques utilized in the automatic image description generation is given. From this survey, the author concludes that the automatic generation of image description can provide a high-quality description of the scene, like humans when compared to the traditional image annotation based on keywords.

The generation of language descriptions of videos with thousands of frames in a few words is demonstrated in [22]. This method uses sparse object stitching and latent topics. The generation of image description can be possible either by top-down approach, that is, generation of language descriptions using the combined form of language models and object detections, or by bottom-up approach, that is, the use of keywords from training to testing images via the nearest neighbor technique. But the use of natural language in generation of image description is still under research. Hence, in this paper, both top-down and bottom-up approaches are combined to generate the description of a video in natural language by considering the most relevant contents of the video. It is achieved through the use of a hybrid model which performs initial keyword annotation in the low level using multimodal latent topic, then concept detection in the middle level, and finally generating language description in the high level. The results obtained from this hybrid model were found to be having a higher agreement with the human descriptions.

The challenges faced when developing a model for performing actions, such as image-to-image search, image-to-tag search, and tag-to-image search through internet images and their accompanying tags, have been explained in [23]. This paper models the image statistics and its related text for internet images to facilitate their retrieval in several ways such as automatic image annotation, image search based on keywords, and image search based on similarities. This can be achieved only when the developed model is accurate, scalable, and flexible. Hence, this system utilizes a 3-view Canonical Correlation Analysis (CCA) model that includes high-level semantic information in its final stage. The 2-view CCA uses the vision and text data and maximizes the correlation between them to handle multiple searching ways of image in the same way. Hence, various image classes are mixed while retrieving it. But the third view represented in this paper utilizes the semantic information, which separated the classes better than the 2-view approach, and provides a significant increase in the image retrieval accuracy for diverse datasets.

A method to generate image descriptions using deep visual semantic alignment has been demonstrated in [24]. The dataset used in this method consists of images and their descriptions in text format. The developed model utilizes Convolution Neural Network (CNN) for image detection, bi-directional Recurrent Neural Network (RNN) for sentence extraction, and finally a structured objective to align the above two into a multimodal embedding. Then, a multimodal RNN is utilized for making the inferred alignments to learn generating the region description. This alignment model is evaluated on 3 datasets, namely, Flickr8K, Flickr30K, and MSCOCO, and the results outperformed the retrieval baselines.

Prediction of similarities in words using word embeddings based on symmetric patterns is described in [25]. The symmetric patterns are automatically obtained by the plain text from a large corpus, in which the coordinates represent the concurrence in patterns of the given word with another word in the vocabulary. This method can be well-suited for predicting the similarities in words, since this system is based on symmetric word relationship, and the features can be modified based on the application. The similarity score obtained from this method is 0.563 higher than the score obtained by word2vec.

A human-robot knowledge transfer framework has been demonstrated in [26]. This paper constructs a robot that learns in real-time human implementations and transfers the knowledge. This robot system acquires, represents, and transfers knowledge. The knowledge has been characterized in Spatial, Temporal, and Causal And-Or Graph hierarchical network. It is considered stochastic grammar. The learning is done continuously online along with the inferences. Here, the robot acts like a knowledge database in which humans feed and retrieve the skills that can be useful for both robots and humans.

The semantic representation model based on visual attributes is described in [27]. The dataset used in this method consists of 500 concepts of the visual attributes and their related 688K images. Instead of image features, this method utilizes visual attributes, since the attributes are not confined to categories and can easily differentiate the concepts more clearly. The dataset is trained on an attribute classifier, and the prediction outcomes are integrated with the textual distributional model for grounding the meaning of the words. The results obtained from this method demonstrate that this bimodal scheme is better than the modal scheme in predicting the meanings of the words represented by humans.

A grounded compositional semantics for generating image description has been implemented in [28]. The RNN-based approach makes use of constituency trees and provides image description by generating feature vectors, but it does not accurately describe the grounding meaning. Hence, this paper introduces a new model called DT-RNN which makes use of the dependency trees for embedding the

sentences into vector space to retrieve the images from the generated text descriptions. This method focuses on the agents and actions in a sentence. Hence, it can generate the description based on the word order and syntactic expressions, and the results obtained outperform the RNN, CCA, and Bag-of-Words approaches.

Book2Movie, a system that aligns the scenes in a video with the book chapters, has been implemented in [29]. The films that are taken from novels have some scenes that are entirely different. This Book2Movie method finds the differences between the source and the adaptations and generates a description of the video from the book. This study uses a dataset obtained from two novels named *Game of Thrones* and *Harry Potter and the Sorcerer's Stone* and their corresponding videos such as 9hrs video for *Game of Thrones* and 2hr 30mins video of *Harry Potter and the Sorcerer's Stone*. The scene detection is performed using dynamic programming, the dialog parsing in the film is conducted by extracting the subtitles in the DVD, and the dialog parsing in the novel is carried out through a hierarchical method for text processing. Finally, a graph-based alignment approach is utilized for aligning the scenes in the video with the book chapters [30-32].

4 Proposed Methodology

Figure 3 shows the conceptual framework of image or video captioning that uses language modeling as another top layer or simultaneous vision and language modeling with the help of specific algorithm or loss function. Unlike the traditional system, the proposed system uses structural multimodal input and also generates structural multimodal output.

In computer vision, a multimedia file can be described by a sentence, or a set of sentences called discourse that explains the complete story behind it. The sentences are first learned through the use of the web-scale corpora to detect the objects, actions, and places related to the image/video. It interrelates the image/video and the natural language texts using computer vision and natural language semantics. The tasks involved in computer vision are tracking, object detection, and event recognition. These tasks are performed concurrently and generate the best probable text description by concentrating on the major events in the image/video for activity recognition.

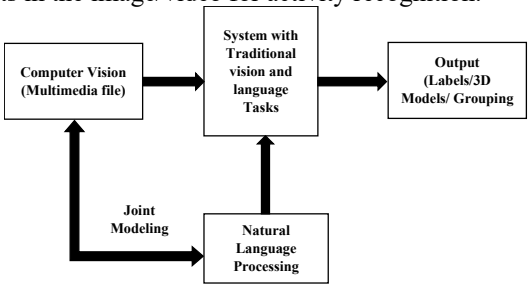

Fig. 3: Conceptual framework.

In NLP, lexical semantics are used to track the sentences. These sentences consist of details, including, "Who did what to whom?", "Where they did?", and "How they did it?". The natural language texts are represented with a set of predefined grammar and predefined vocabulary, which are given as follows: a noun phrase is used to describe an object, a verb is used to describe an action, an adjective is used to describe a noun, and an adverb/a preposition is used to describe the event characteristics.

The outputs from the computer vision model are given to the NLP model and are converted into speech output to assist the blind users.

3.1 System Design

The proposed system uses a smartphone for navigation and speech guidance. The smartphone records the blind's surrounding areas continuously, processes them, and performs object recognition to inform them about the objects near them in natural language text which is converted into speech output by the smartphone. Faster R-CNN algorithm is utilized for performing the computer vision-based object recognition task. Then, the Bag-of-Words approach is used to convert the text into natural language, and this text data is given to the server. Finally, the text-to-speech action is performed by the smartphone to guide the blind with speech output. The overall design of the proposed system is shown in fig. 4.

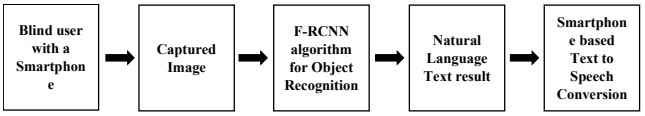

Fig. 4: Overall design of the proposed system.

The Faster R-CNN algorithm consists of 2 modules for object detection, namely, the Region Proposal Network (RPN) which is a deep fully convolution network module to propose the regions that are fed into the next module called Fast R-CNN detector. Figure 5 illustrates the structure of F-RCNN.

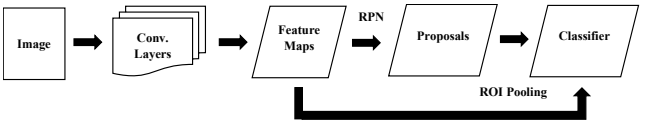

Fig. 5: Structure of F-RCNN.

The RPN modules induce the F-RCNN to detect the objects in an image. After the image is processed, the image with several convolutional layers and the max-pool layers produces the feature maps. The RPN provides a set of object proposals with object scores. A small crop of the image is chosen as the Region of Interest (ROI) which produces a vector with fixed-length features obtained from the feature maps. It is then fed into the classifier by ROI pooling. The rectangular proposals from RPN and the vectors are combined for final classification.

This object recognition output from the vision task is given as the input for natural language processing to convert the text into natural languages, which can be easily

understandable by humans. Before converting the text to natural language, it is first to be cleaned to save the memory space and processing time, while converting the text into vectors. The text cleaning is done by removing the HTML tags like
, <h1>, <p>, and so forth, if available, since they do not return any valuable information needed for the users but increase the processing time and storage consumption. Then, the punctuations are removed, since these too are similar to the HTML tags. Next, stop words are removed, such as 'is', 'that', 'there', 'this', and so forth, since they do not return any valuable information but create clutter in the storage, and then stemming is performed, which means the extraction of the root words; for example, the words 'hopeful' and 'hopefully' give the same meaning as 'hope', which is the root word. If vectors are created for all the variations of the words, they are unwanted, since all stand for the same meaning, and finally all the texts are converted to lower case, since there is no need of creating vectors for words like 'Chocolate' and 'Chocolates' since both are the same; only one vector is enough. If words are not converted to lower case, then it ends up creating vectors for both. Moreover, it is important to ensure that there are no words in alphanumeric.

After cleaning the words, they can be subjected to vector conversion to make them understandable for the machine learning algorithms that are used to convert the text into natural languages via NLP. In this paper, Bag-of-Words (BOG) approach is used for natural language processing. NLP works only with numbers; hence, the text data cannot be fed directly into the algorithm. BOG algorithm converts the text into a bag of words which stores the count of the most frequent word occurrences. The most frequent word can be found by declaring a dictionary to hold the bag of words; then, the sentences are tokenized into words, and these words are checked in the dictionary for their existence. If the word exists in the dictionary, the count will get increased by 1; otherwise, this word will be added to the dictionary and set the count as 1.

After converting the texts into natural languages, which is an NLP task, it is then converted into speech so that it can be easily understood by blind users. This can be performed by sending the natural language text to the smartphone via a server, and the smartphone performs the translation task and provides the speech output to the user.

5. Experiment

5.1 Various Experimental Tasks

5.1.1 Sentiment Classifications (SentiC)

The method Sentiment Treebank (STT) [33] is a database forecasts the favorable or negative sentiments of film reviews. The provided mix of 6930 train, 855 development, and 1675 test phrases will be used. Labeled phrases that appear as actual requirement of training sentences are treated as separate training occurrences. Accuracy in measurement.

5.1.2 Relation Classification (RC)

It comprises of 8000 words in train and 2700 texts in test that were individually tagged with 20 connections (15 focused relations and Even other) [34]. We utilize 1500 training instances as developed since there is no develop set, comparable to F-Measure.

5.1.3 Textual Entailment (TE)

The premise-hypothesis pairings in Natural Language Inference (SNLI) [35] are labeled with such a relationship (entailment, , neutral ,contradiction). We have 549,400 pairings for train, 9,840for develop, and 9,850 for testing after deleting unnamed pairs. Measurement: precision.

5.1.4 Answer Selection (AS)

A database of open-domain question-and-answer questions employ the specific task that assumes each question has at least one right answer. The associated dataset contains 20,355 issue candidate pairings in train, 1,134 in development, and 2,390 in testing, with the conventional configuration of only evaluating questions with right answers in testing [36]. The objective is to select the proper outcome(s) for an inquiry from a candidate list. MAP and MRR are two indicators.

5.1.5 Question Relation Match (QRM)

We use the WebQSP [37] datasets to make a large relation identification task, taking use of the labeled semantics parses of questions that are available. Subject entities from the parsed; (ii) selected all the connections chains (length 2) linking to the main entity; and (iii) define the connections in the labeled parse as positives and the others as negatives for each concern. This challenge may be rephrased as a sequential matching issue. For training, ranked team is employed accuracy in measurement.

5.1.6 Path Query Answering (PQA)

PQA contains KB paths such as $(e_h, r_{0,} r_1, \dots\dots\dots r_t, e_t)$, where e_h is a head entity, $r_{0,} r_1, \dots\dots\dots r_t$ are relation sequence, e_t is tail entity. 6,266,058/27,163/109,557 are the paths for train/ develop/ test.

5.1.7 Part of Speech Tagging

Part-of-speech labelling (POS tags, PoS tagging, or POST), also known as grammar tagging, is the act of marking up a word in a text (corpus) as relating to a certain part - of - speech, depending on both its meaning and context.

5.2 Experimental Setup

Our tests are set up in the following way to fairly examine the encoding capabilities of several basic DNNs are always start from the beginning, with no prior information, such as word embedding. (ii) Always train with a simple configuration that does not include advanced techniques like batch normalization. (iii) Find the best hyper - parameters for each job and model individually, such that all outcomes are based on the best hyperparameters. (iv) Examine each model's fundamental architecture and application: The convolution layer and the max-pooling layer are the two layers that make up CNN. GRU and LSTM analyze the inputs from left to right, usually using the most recent hidden layer as the final approximation and additionally include bi-directional RNNs for POS tagging, as this ensures so each term's representation may include the word's meaning on both sides, as the CNN provides. The Best results of CNN, GRU and LSTM tasks in Language Processing as shown in table 1.

Table 1: Result Comparison for CNN, GRU and LSTM using NLP task.

	Methods		Performance	lr	hidden	batch	SentLen	Filter_size	margin
Texi Classification	SentiC (Accuracy)	CNN	89.02	0.5	35	19	60	2	0.02
		GRU	81.05	0.3	25	20	60	1	0.01
		LSTM	75.24	0.2	15	5	50	1	0.01
	RelationC (FI)	CNN	75.05	0.5	45	19	40	3	0.5
		GRU	65.23	0.2	35	27	50	1	0.3
		LSTM	56.43	0.1	22	15	60	1	0.3
S-Match	Textual Entailment (accuracy)	CNN	90.02	0.6	55	65	20	2	0.02
		GRU	83.05	0.4	35	92	50	2	0.01
		LSTM	73.22	0.2	15	80	30	1	0.01
	Answer Selection (MAP&	CNN	69.02	1.5	32	45	30	3	0.025
		GRU	55.93	1.1	24	15	40	2	0.22
		LSTM	45.02	0.8	12	76	50	2	0.01
	QRM (accuracy)	CNN	88.93	2.4	75	45	10	3	0.4
		GRU	75.80	1.5	45	50	20	1	0.4
		LSTM	60.02	0.8	33	35	10	1	0.3
Sequence Order	Path Query Answering	CNN	58.09	0.4	85	55	50	5	0.2
		GRU	50.20	0.3	73	32	30	1	0.2
		LSTM	45.02	0.1	65	54	40	1	0.1
Context Develop	Part of Speech (accuracy)	CNN	98.56	5.8	51	42	20	4	0.01
		GRU	85.34	2.3	45	72	10	2	0.02
		LSTM	70.34	2.0	32	27	30	1	0.01

Hiding size, reduced sample size, number of epochs, maximal phrase length, frame size (for CNN only), and margins in rank loss in AS, QRM, and PQA activities are all tuned on development.

5.3 Result and Quality Analysis

Table 1 provides the research observations for all assignments and models, as well as the parameters that go with them. GRU outperforms SentiC in TextC and is comparable to CNN in RC. CNN beats GRU (and also LSTM) in SemMatch on AS and QRM, whereas GRU (and also LSTM) beats CNN on TE. Both GRU and LSTM outperform CNN in SeqOrder (PQA). CNN beats each RNNs in ContextDep (POS tagging), but falls short of bi-directional RNNs. The SeqOrder and ContextDep outputs are as expected: Information or assistance (for PQA) and long-range context dependence are well-suited to RNNs (for POS tagging). However, there are some surprise findings in the other two main categories, TextC and SemMatch. CNNs are thought to be strong at identifying local and stance features, thus they should do well on TextC; nevertheless, in our tests, RNNs outperformed CNNs, particularly in SentiC as shown fig. 6. Recurrent neural networks can encode the entire input's structure dependent interpretation.On SentiC, we perform some failure analysis to explore the surprising observations.

When it comes to Quality Analysis, CNN works best once the sentiment is decided by the full statement or long-range semantics dependence – rather than just a few local key-words. Example (1) has the terms "won't" and "miss," which are often associated with negative emotion, but the entire sentence expresses a positive attitude; hence, a brief architecture such as GRU is required. From the other hand, modeling the entire phrase can be time consuming and might lead to the omission of important details.

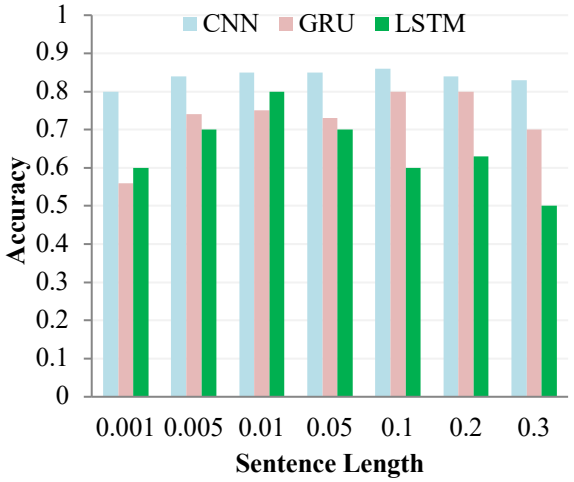

Fig. 6: Distributions of sentence length and accuracy

The GRU captures the full long sample (3) word sequence, finding it challenging for the negative key to play a significant role in the ultimate depiction. Example (4) appears to be positive in the first section but negative in the second. Because GRU uses the latest hidden layer to describe the phrase, it is possible that the predictions will be incorrect. Examining accuracy and long sentences can also help. Figure 7 indicates that sentence lengths in SST are predominantly short in train, whereas develop and test neither are near to nor mal dispersion around 20. The accuracies in length ranges are depicted in fig. 7 discovered that when sentence lengths are modest, such as 10, CNN and GRU are equivalent; nevertheless, when larger phrases are encountered, GRU gains a growing edge over CNN. Long sentences in SST are generally made up of inverted semantic clauses, such as "this edition is not classic like its predecessors, but its delights are still ample," according to error analysis.

6 Discussion

In this research examined Computer vision and Natural Language Processing and based on multimedia and robotics integration. As shown in visual characteristics, image captioning, video captioning, and graphics, natural language processing provides high-level context to aid low-level machine vision operations in big corpora. Machine translation in robotics can assist a robot in performing more accurate thinking and management given the input stream generated by low-level computer vision techniques [38]. The overarching subject of distributional semantics as a notion capable of designing a product for computer vision applications processing is then highlighted.

We looked at both old methods like bag-of-words systems and feature based, and more current methods like word2vec and pattern recognition.

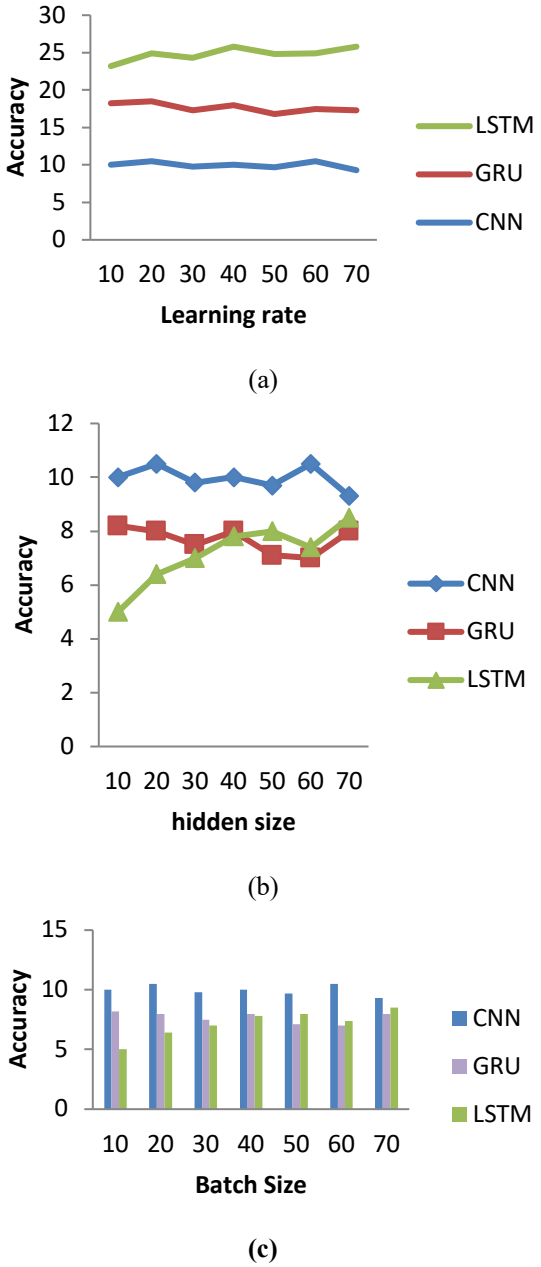

Fig. 7: Accuracy Comparison between Learning rate, Hidden size and Batch Size.

Finally, we compile all of the techniques into a redistributive semantics framework and offer a glimpse into a current cognitive-inspired model. The present status of the discipline shows rapid advancement that may be summed up in three words: accuracy, scalability, and innovation. Deep learning's widespread success is responsible for the accuracy. The breakthroughs in Hardware acceleration of high-performance computation enable the scalability. Many unique implementations, such as captions, voice recognition, and conversation systems, demonstrate the ingenuity. Previous research difficulties

have made significant progress, and the discipline has progressed to the next level, always pushing the boundaries [39]. Deep learning using CNNs outperforms standard techniques in computer vision in terms of accuracy, as well as the system can be taught end-to-end, through input to output. Even yet, there are obstacles lurking on the edges of the system's capabilities. For example, a system's ability to categories a large lot of items is still far from visual acuity.

Furthermore, it is still far from lead to more effective to detect accurate organized knowledge and understanding of fine-grained categorization. When a person sees a bird, for examples, he has a lot of background knowledge gained up when he hears a bird. He must also confirm that the bird is flying or has the capacity to fly (recognizing that the bird does not have a damaged wing or that it is still alive is critical). A bird is a sensory stimulation, and his mind considers what it has seen in the context of past knowledge. That is something that the existing framework is incapable of doing effectively. To execute a scaled structural prediction, at the very least, a new model architecture is required. Learning based with RNNs, embedding models like embedding, and memory models like LSTM outperform classical techniques in natural language processing in terms of sequences to organize learning accuracy. However, when the structure of the input data is unknown ahead of time, LSTM might fail. In such circumstances, an extension likes TreeLSTM [40]. which handles forest data, or even fodder neural networks with weight matrix, which handles tree-structured data, may provide a superior outcome.

This research presents a new topic to consider, particularly for a system that includes an attention function. Without a thorough understanding of technique (a superb set of tools), no system can function successfully. Modeling bidirectional structured data using tensor decomposition is a promising future area. In the same way that SVD decomposes a tensor into a collection of orthogonal bases, we may deconstruct a tensor into relevant statistics. This might be a suitable starting point for next iteration of multimodal distribution of income semantics meaning encoding. A tensor may represent a wide range of information. For topic models, we may create a tensor of word co-occurrences. To do graph segmentation, we can generate a tensor of connected data. The three main extensively used DNNs – CNN, GRU, and LSTM – were compared in a sample group of NLP tasks in this study. Except when the job basically works effectively such as in some sentimental analysis and question-answer answering settings, we discovered that RNNs perform well and are resilient in a wide range of tasks. Furthermore, concealed size and batch size can have a significant impact on DNN performance. This shows that optimising these two parameters is critical for excellent CNN and RNN performance.

7 Conclusions

The current study intends to provide a solution for the problems face d by blind or visually impaired people for their movements automatically and without anyone's help. The blinds and visually impaired experience lots of trouble while walking through a busy road or to a place that they are unfamiliar with. This smartphone-based guidance model solves the blind and visually impaired people's problems related to their navigation from the start to the destination and helps them in avoiding obstacles and makes them aware of their surroundings for smooth movements. The proposed model is a simple, user-friendly, and low-cost design which can improve the lifestyle of the visually impaired and blind people by guiding them throughout their way for their mobility and independent life.

In this paper, two AI-related tasks, that is, computer vision and natural language processing, in the field of multimedia and robotics have been discussed and applied on a particular task like recognizing the objects in an image through computer vision, converting the text output of the recognition task into natural language through the use of natural language processing, and finally converting the natural language information into speech for guiding the movements of the blinds and visually impaired people.

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 (ICoDT2)*, 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, Book2movie: 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. ArXiv150805326*, 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. ArXiv160601847*, 2016.

[39] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *ArXiv Prepr. ArXiv14042188*, 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.