Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 3
Open Access

Mythological Story Teller: Extending Scope Of Deep Learning To Indian Mythology

Aditya Goutam[1] , Sunit Trivedi[2] ,Veeraj Goudar[3] , Nehaal Pandey[4] , Sunita S. Barve[5]
[1,2,3,4,5]MIT Academy of Engineering, Pune, India
[1]aditya.goutam@mitaoe.ac.in, [2]trivedi.sunit@mitaoe.ac.in, [3]veeraj.goudar@mitaoe.ac.in, [4]nehaal.pandey@mitaoe.ac.in, [5]sbarve@mitaoe.ac.in

Cite this paper as: Aditya Goutam, Sunit Trivedi,Veeraj Goudar, Nehaal Pandey, Sunita S. Barve (2024) Mythological Story Teller: Extending Scope Of Deep Learning To Indian Mythology. *Frontiers in Health Informatics, 13* (3), 4749-4763

A B S T R A C T

Mythology holds immense importance as it serves as a repository of ancient wisdom, moral lessons, and cultural values. At the same time, Mythological Scene Recognition presents a unique challenge in Computer Vision as it demands precise identification and interpretation of characters and their relationship. Also, the lack of a dataset for detecting and identifying Mythological characters poses a significant obstacle. Subsequently, this paper presents a novel approach to Mythological Scene Recognition utilizing an integrated framework involving image processing, image recognition, web-based scrapping and a part of text analysis for summarization. The Mythological Scene Teller algorithm is proposed, which would fetch a picture of the past, identify the characters present in the scene, and give a summarized text representing the scene. A comparative study with different character recognition and classification models is also presented, treating the characters as objects. The paper incorporates a robust and efficient Web-scraping model to web scrape available information about the characters and establish relationships between the characters along with the integration of T5 Base for text summarization and information retrieval culminating an effective and enhanced approach.

Keywords: *Computer Vision; Character recognition; Deep Learning; Information Retrieval; MST (Mythological Scene Teller) algorithm; Mythological Scene Recognition; Text Summarization; Web-Scraping.*

INTRODUCTION

The Indian mythology consists of a rich and complex set of characters. These characters and their stories have been immortalized in the form of sculptures, temple carvings, paintings, and other art forms across the entire subcontinent (A. R. Bharadwaj et al., 2020). The epics of the Ramayana and Mahabharata are just two examples of the vast tapestry of stories that make up Indian mythology. The significance of this mythology extends beyond mere art and entertainment, as it has a profound influence on the values, ethics, and spiritual growth of millions of people. However, in our fast paced, technology- driven world, there is a growing gap between the younger generation and this invaluable wisdom. Modern distractions and changing lifestyles have led to a disconnection from these narratives, posing a risk of losing touch with essential cultural and moral values. To address this gap, there is a need for innovative approaches that can bridge the generational divide and reconnect the younger audience with the wisdom embedded in mythology. One of the significant challenges in bridging the gap between the younger generation and Indian mythology is the lack of accessibility and relatability. The younger population, raised in a digital-centric world, often finds it difficult to engage with traditional forms of art and narratives. Recognizing and understanding complex mythological scenes in paintings and sculptures can be a daunting task, and this process is further hindered by the scarcity of accessible and concise information. The dearth of easily available resources for identifying the characters, stories, and teachings depicted in these artworks compounds the challenge. In the absence of such resources, the younger generation is at risk of missing out on the profound wisdom and cultural identity that Indian mythology imparts.

Figure 1. Input image and Output Image

The proposed Mythological Story Teller (MST) when incorporated into an application, will take in a painting of a mythological era as an input, pre-process the image, and identify classes i.e. characters and scenes, and establish relationships between them to scrape data about the scene and provide a summarized paragraph about the details of the image. (Figure 1) shows an example of an input image of a mythological scene of the Mahabharata, Indian mythology and the output image which includes characters recognized by the MST algorithm. The classes detected are ['Draupadi', 'Dusashana', 'Krishna', 'Vastraharan']. To overcome the difficulty of scene recognition in mythology, this research tried a novel approach by defining major events of the past as a class. (Vastraharan in the above example is a scene detected as a class.) This approach helped improve the MST algorithm overall.

Previous attempts at recognizing mythological scenes faced significant challenges. The main difficulty arose from a shortage of available sets of data featuring sculptures specifically depicting Indian mythology, slowing down the development of strong recognition models. In addition, previous implementations mainly focused on general scene recognition, lacking the necessary customization for understanding the detailed subtleties of mythological representations. This absence of specialization greatly hindered the effectiveness of these efforts. Moreover, many earlier approaches did not fully use 2 advanced deep-learning neural networks, a crucial element in achieving higher accuracy and reliability in recognition tasks. To address these challenges, the proposed Mythological Story Teller (MST) algorithm aims to leverage cutting-edge technology to make Indian mythology more accessible and engaging for the younger generation. The Mythological Story Teller (MST) algorithm addresses the challenges in recognizing complex mythological scenes in paintings and sculptures, allowing for seamless identification of images by users.

The key features of the MST algorithm lie in its adeptness at establishing relationships between the identified characters and scraping relevant information available on the internet. The algorithm's pivotal significance lies in its capacity to identify the scene present in a painting and provide a summarized overview of the same. The primary contributions of this research paper are as follows:

1. Adapting to the intricate, often stochastic nature of mythological scenes.

2. Employing character detection model for comprehensive object and component detection within the scene.

3. Establishing relationships between identified characters for extended context.

4. Integration of Web-scraping capabilities for acquiring information about characters and their relationships from reliable web sources.

5. Incorporation of T5 Base for text summarization and information retrieval, alongside avoiding duplicate text for a detailed summarized gist of the scene.

The subsequent sections of this journal are organized as follows: Section 2 provides an overview of existing approaches in mythological scene recognition. Section 3 provides the system architecture of the MST algorithm. In Section 4, mathematical modelling is covered. Section 5 covers the MST algorithm and its underlying methodology. Section 6 encompasses performance analysis, featuring the evaluation and results of the MST algorithm. Finally, Section 7 outlines the results achieved from the research, and Section 8 concludes the research.

1. LITERATURE REVIEW

The research comprises completing a full literature assessment on scene recognition techniques, including both traditional and cutting-edge procedures. It entails reviewing scholarly publications that focus on character recognition and character connection inference using the YOLO (You Only Look Once) paradigm. Furthermore, the inquiry includes a review of research publications describing the T5 Base model's applicability in the specified subject.

2.1 Scene Recognition using traditional methods and advanced methods

Scene recognition is one of the most important and integral part of mythological scene recognition. The majority of cutting-edge visual identification techniques are created using general-purpose datasets and ignore the uniqueness of scene data. The challenge of scene identification is addressed in this study by the efficient Scale Attentive (SA) module, which streamlines the scale-aware attention learning pipeline to support the feature re-calibration and refining process (X. Yuan et al., 2022). This paper (L. Liu et al., 2020) outlines a technique for automatically identifying episodes from Indian mythology in paintings and line drawings. To assist with scene detection, artificial neural networks were utilized to recognize legendary creatures, animals, scenery, and weapons in the input image. The mythical writings linked to the image were utilized to evaluate the character relationships and create a Character Association Graph, which was then used to enhance the neural network predictions. In this study, (B. Labinghisa & D. M. Lee, 2021) have employed deep convolutional neural networks to enhance ImageNet, a dataset for object detection used to train a scene dataset that can identify indoor environments in academic institutions. The proposed scene identification technique is evaluated with various models trained in Places365 to compare which performs better for a new dataset specialized in indoor space. High accuracy is required for scene recognition applications in indoor environments. Using scene attributes as supplementary features to object and scene characteristics, the paper discusses the discrimination of scene attributes in local regions in this study (H. Zeng et al., 2020). These visual features are extracted from two distinct CNN branches, one of which extracts the image's global features and the other of which extracts the features of its local areas. The Places Database is a collection of 10 million scene images that have been classified using scene semantic terms. (B. Zhou et al., 2020) present scene classification CNNs (Places-CNNs) that greatly outperform the prior methods using cutting-edge Convolutional Neural Networks (CNNs). The Places Database presents a fresh resource to direct future progress on scene recognition difficulties because of its vast coverage and high diversity of exemplars. The goal of this work is to provide a review of the state-of-the-art in deep learning models for scene detection from visual data. Scene recognition is a still developing area of computer vision studied from both a static and dynamic image standpoint. This paper (B. Oh et al., 2018) discusses the different datasets available for scene recognition and various ensemble techniques. To create a discriminative local semantic representation, (Q. Bi et al., 2021) present in this paper a local semantic enhanced ConvNet (LSE-Net) for aerial scene recognition that mimics human visual perception of key local regions in aerial scenes. A local semantic perception module, a classification layer, and a context-enhanced convolutional feature extractor make up our LSE-Net. In this study, (H. Seong et al., 2019) propose a score level Class Conversion Matrix (CCM) with a strong emphasis on the relationship between objects and scenes for scene recognition. Many existing techniques have previously developed scene identification algorithms that take into account the intimate connection between objects and scenes. The majority of these algorithms, however, do not do any conversions or reconstructions before using the object attributes, therefore it is unclear whether they are effective at accurately identifying scenes. In this paper, (Z. Wang & Z. Li et al., 2020) discuss the advent of computer graphics. Computers are being developed with the ability to swiftly find and recognize text in photographs of natural surroundings. To accomplish their goal, they examine the computer graphics (CG) tool. The branch of mathematics transforms 3D pictures into several types of grids that may be seen on a computer display. It usually refers to the efficient management and dissemination of multimedia information through computer management or, later, local information systems. the optimization of the core network structure lowers the impact of situations on object detection. A technique called DFCRF (Deep Forest with Convolutional Residual Features) is suggested in this research (M. Han et al., 2018) and the authors take advantage of the recently made AI Challenger dataset, which has only about 50,000 photos and was primarily taken in China. Instead of using simply CNNs, the authors combine gradient-based XGBoost and cascade deep forest with convolutional residual features for additional recognition. Then, to demonstrate the efficacy of their approach, they run the dataset and the reconstructed Places2 dataset. This study (Z. Qu et al., 2016) provides a color fusion approach based on a combination of scene classification, fusion quality measure, and color transfer. They use the gist descriptor and SVM classifier-based scene categorization technique in the color fusion algorithm. For each classified input image, they employ the suggested color fusion quality measure structure to determine which reference image is most closely matched. While applying the color transfer method, they obtain a high-quality color fusion image. In (T. Zheng et al., 2020), a spatial-temporal attention mechanism is employed to increase the precision of picture recognition to enhance the model's performance. The completely automated coal face, the route leading to the coal mine, and the mining equipment are among

the three sorts of model sample sceneries featured in this research.

2.2 Character & Scene recognition

YOLO is used for character recognition in mythological scene recognition. In this paper (C. Liu et al., 2018), YOLO is a single-stage CNN-based algorithm for object detection and classification and is one of the quickest objects detecting algorithms with high accuracy and excellent real time performance. (W. Fang et al., 2020) proposes Tinier-YOLO, a compact variant of the YOLO (You Only Look Once) object detection model, aimed at running efficiently on embedded devices. While previous versions like TinyYOLO-V3 were designed for such constrained environments, Tinier-YOLO further reduces model size while improving detection accuracy and real-time performance. It achieves this by incorporating the fire module from SqueezeNet, optimizing the connectivity between fire modules, and introducing a passthrough layer for fine-grained feature merging. Tinier-YOLO achieves competitive performance in terms of mean average precision (mAP) on datasets like PASCAL VOC and COCO when compared to other lightweight models. (S. Li et al., 2021) proposes a smaller version of YOLO, Fast YOLO, which achieves an astonishing 155 frames per second while maintaining double the mean average precision (mAP) of other real-time detectors. It formulates object detection as a regression problem, predicting bounding boxes and class probabilities directly from full images. This unified architecture enables end-to-end optimization for detection performance. (J. Tao et al., 2021) discuss the research on the vgg16 convolutional neural network feature classification algorithm based on transfer learning. (M. Yavartanoo et al., 2021) proposes a DNN-based method (PolyNet) and a specific polygon mesh representation (PolyShape) with a multi-resolution structure for 3D shape recognition. (P. C. Blaud et al., 2022) deals with model predictive control synthesis which takes benefits from artificial neural networks to model (non-linear) dynamical systems. Residual networks (ResNet) and PolyInception networks (PolyNet) neural network architectures are used for image recognition.

2.3 Text Summarization Models

T5 and GPT-2 models are used for text summarization. (M. R. Suryakusuma et al., 2023) proposed work which examines the English-to-German machine translation performance of the Text-to-Text Transfer Transformer (T5) paradigm. The study's objectives are to assess the quality of the dataset and tokenizer format, compare the effectiveness of various generation techniques, and investigate the benefits and drawbacks of the T5 model. Different scoring systems are provided by BERTScore and BLEU, the evaluation measures that are employed. The outcomes demonstrate the usefulness of the Greedy Inference approach in this challenge, as it receives the top scores in both criteria. The study sheds information on how well the T5 model performs while translating from English to German and emphasizes the need for more advancements in translation accuracy and lexical context. Based on our studies, we found that the Greedy Inference approach consistently produced the best results in terms of BERTScore and BLEU metrics. Generating questions is useful for evaluating reading comprehension, extending datasets for question answering, and eliciting spontaneous queries in chat systems. Numerous models have been employed in earlier research to generate questions from contexts, but none of them worked well in lengthy contexts. To get around this problem, this study, created questions using intermediate context representations like knowledge graphs (K. Aigo et al., 2021). In this work, they concentrated on creating questions utilizing the T5 language model and knowledge graphs. Trained the model by explicitly maintaining the graph's structure, and utilized the language model to generate questions based on the knowledge graph. The bidirectional Graph2Seq model (G2S) and the T5 language model, both with and without mask, were equivalent as a result of the automated evaluation. The purpose of this work is to review the research publications and present an abstract summary of them. BERT models fared well in extractive summarization, however, there is a need for improvement in abstractive summarization. Using the distill-GPT2 version with more computer capabilities, we use regularization to decrease local similarity while concurrently enhancing global similarity (N. Darapaneni et al., 2023). Transliteration-based Generative Pre-trained Transformer 2 (GPT-2) model is proposed in this study to summarize online news articles for the Tamil language by extracting a large number of relevant features such as sentence position, one 4 hot encoding, number of entities, term frequency, and inverse document frequency. The tests are carried out to analyze the improved transliteration model based on the Bilingual Evaluation Understudy Score, as well as the GPT-2 based text summarization model based on the ROUGE evaluation measures (C. R. Dhivyaa et al., 2022).

Web scraping is also used in this research. (Lunn et al., 2020) demonstrate how web-scraping can be useful in extracting data from publicly available web pages and in addition they also discuss how natural language processing can be useful to obtain salient information from textual data. They have demonstrated techniques that can be used for numerous applications. Also, they show how the application of web-scraping and NLP are useful in obtaining and analyzing pertinent information from the internet.

2.4 Relationship Establishment

The research has introduced a new dataset, VRDU (Visual Relationship Detection) which contains images of objects with unknown relationships) and proposed a model MF-URLN (Multi-modal Feature-guided Undetermined Relationship Learning Network) (J. Wu et al., 2021). The paper addresses an important problem in computer vision: detecting relationships between objects in images. This paper focuses on undetermined relationships that are not predefined. The MF-URLN model combines visual appearance features, semantic information, and contextual cues to detect relationships. The multi-modal feature-guided object detector utilizes visual and semantic features to detect objects in the image. This component integrates visual appearance features and semantic information to improve object detection accuracy. (J. Zhu & H. Wang, 2022) introduces a novel approach called the Multiscale Conditional Relationship Graph Network (MCRG-Net) to tackle the challenging task of referring relationship recognition in images. The proposed MCRG-Net aims to address these limitations by incorporating multiscale contextual information and conditional relationships into a unified framework. The MCRG-Net model consists of three main components which are a multiscale feature extraction module that captures contextual information at different scales, and a conditional relationship graph module that models the dependencies between objects based on their appearances and spatial relationships, and a referring relationship recognition module that predicts the referring relationships between objects mentioned in the query.

(A. -A. Liu et al., 2021) introduces a novel approach called Adaptively Clustering-Driven Learning (ACDL) for visual relationship detection. Visual relationship detection involves identifying and understanding the relationships between objects in images, which is a challenging task due to the complex nature of these relationships. The proposed ACDL method addresses the limitations of existing approaches by leveraging adaptively learned clusters to guide the learning process. (H. Shalma & P. Selvaraj et al., 2022) presents a method for occlusion detection in images that takes into account the distance between objects and employs focused attention mechanisms. The proposed approach addresses this challenge by incorporating distance awareness and focused attention mechanisms. The distance-awareness component allows the model to understand the spatial relationships between objects and utilize this information for occlusion detection. By considering the distances between objects, the model can distinguish between occluded and non-occluded instances more accurately. The focused attention mechanism guides the model's attention to relevant regions of the image, enabling it to focus on potentially occluded areas. (X. Wu et al., 2022) presents a method for unsupervised change detection in multimodal remote sensing images. The proposed approach addresses the challenge of unsupervised change detection by leveraging structural relationship graph representation learning. The method constructs a structural relationship graph that captures the spatial relationships between image patches in the pre-change and post-change images. The structural relationship graph is learned through a graph neural network (GNN) framework, which incorporates the spatial and contextual dependencies between patches.

There is a gap in the literature on mythological scene recognition using computer vision algorithms. Previous research has focused on specific mythologies and has not used advanced techniques such as deep learning. There is no research on using natural language processing to detect mythological scenes in text. Further research is needed to explore the use of advanced techniques for detecting mythological scenes in various media formats (images, videos, and text).

2. SYSTEM ARCHITECTURE

This research proposes a system for analyzing and contextualizing historical images, structured into three phases. Phase I involves annotating character and scenic classes to create a dataset for training a custom Object Identification model. Phase II uses this model for precise recognition and web scraping to contextualize identified characters. Phase III employs a summarization model to generate concise summaries, offering insightful narratives of the images' historical contexts.

3.1 Dataset Creation

In the initial phase (Phase I), the dataset is meticulously curated through systematic annotation of character classes and scenic classes. This process facilitates effective categorization and contextualization of image content. The annotated dataset is subsequently exported for training the YOLO model, incorporating customized parameters to meet specific research requirements.

3.2 Object Identification and Classification

Phase II focuses on the identification and classification of objects within historical image instances. An input image,

provided by the user, undergoes meticulous preprocessing. The image is then subjected to a custom-trained YOLO model capable of character classification and scenic class recognition. This phase centers around the identification of characters and their associated relationships, utilizing the dataset dictionary as a reference. To enhance contextual understanding, character relationships and identified classes are leveraged in a web scraping process. This operation extracts information from authoritative sources, such as Wikipedia, to pinpoint historical instances in which all identified characters coexisted, thus contextualizing the image content comprehensively. (Figure 2) shows the system architecture of the whole MST Algorithm.

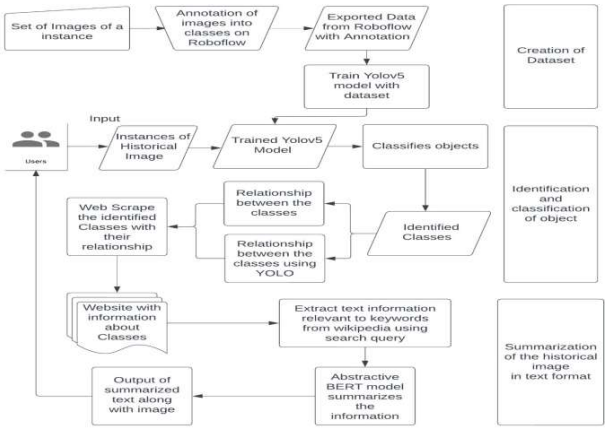


Figure 2. System Architecture

3.3 Summarization of Historical Image Instances

The third and final phase (Phase III) is centered on the transformation of extracted data into a concise format. Information acquired through web scraping, in conjunction with character relationships, undergoes a tokenization process. Subsequently, a T5-based BERT model is employed to generate concise textual summaries. These summaries distill the textual content associated with the image into three to four sentences, presenting a brief yet insightful narrative of the historical context.

3. MATHEMATICAL MODELING

This research presents mathematical models for mythological character recognition using CNN and Transformer models. In YOLO, class probabilities are predicted using softmax, bounding box coordinates are estimated with sigmoid and exponential functions, and confidence scores are calculated with a sigmoid function. FRCNN's region proposal network generates bounding box candidates with objectness scores, and bounding box coordinates are predicted relative to anchor boxes. The Transformer model's multi-headed attention mechanism relies on scaled dot-product attention.

4.1 Equations of Mythological Character Detection Model

YOLO forecasts the class probabilities for the objects contained in each bounding box.

$$\boldsymbol{P(Class \mid i,j,b)} = \frac{e^{z_{i,j,b}^{class}}}{\sum_{c=1}^{c} e^{z_{i,j,b}^{c}}} \qquad (1)$$

where, $\boldsymbol{P(Class \mid i, j, b)}$ is the probability that the object in bounding box $\boldsymbol{b}$ at grid cell belongs to class, and $\boldsymbol{z_{i,j,b}^{class}}$ is the raw score for class at grid cell and bounding box $\boldsymbol{b}$.

YOLO estimates the bounding box's coordinates in relation to the grid cell it is in. These coordinates correspond to the box's width and height (*w, h*) as well as its center (*x, y*). These coordinates are estimated in relation to the grid cell's dimensions.

$$\boldsymbol{x = \sigma(z_{i,j,b}^{x}) + i} \qquad \boldsymbol{y = \sigma\left(z_{i,j,b}^{y}\right) + j} \tag{2}$$

$$\boldsymbol{w = e^{z_{i,j,b}^{w}}}. P_w$$

$$\boldsymbol{h = e^{z_{i,j,b}^{h}}}. P_h$$

Here, $\boldsymbol{\sigma}$ represents the sigmoid function, P_w and P_h are anchor box widths and heights.

YOLO forecasts a confidence score (*Conf*) for every bounding box, signifying the model's assurance on the existence of an object within the box.

$$\boldsymbol{Conf = \sigma\left(z_{i,j,b}^{Conf}\right)} \tag{3}$$

Here, $\boldsymbol{\sigma}$ is the sigmoid function.

4.2 Equations for Faster R-CNN (FRCNN)

Bounding box candidates, or region proposals, that are likely to include objects are produced by the Region Proposal Network (RPN). Here *z* is raw objectness score.

$$\boldsymbol{Objectness\ Score = \frac{1}{1+e^{-z}}} \tag{4}$$

The bounding box's coordinates (*x, y, w, h*) in relation to the anchor box are predicted by the object detection head.

$$\boldsymbol{x = t_x \cdot width(anchor) + center(anchor)}$$

$$\boldsymbol{y = t_y \cdot height(anchor) + center(anchor)} \tag{5}$$

$$\boldsymbol{w = exp(t_w) \cdot width(anchor)}$$

$$\boldsymbol{h = exp(t_h) \cdot height(anchor)}$$

4.3 Equations for Transformer model used

The Multi-head Attention, which is a crucial component of the Transformer encoder and decoder, depends on the scaled dot product attention.

$$\boldsymbol{Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V} \quad (6)$$

Q represents the query matrix. *K* represents the key matrix. *V* represents the value matrix. $\sqrt{\boldsymbol{d_k}}$ is the dimension of the key vectors. The *SoftMax* function normalizes the attention scores.

4. METHODOLOGY

The MST (Mythological Scene Teller) algorithm has been divided and implemented into four phases:

5.1 Visual Perception

Visual Perception is responsible for the initial processing of the image, which serves as the foundation for the Mythological Scene Teller. To accomplish this, the use of a very popular concept is used, that is widely seen in computer vision libraries renowned for its ability to facilitate image input and manipulation. Visual Perception, as the name insists, streamlines the process of acquiring the image and preparing it for subsequent analytical and recognition tasks. These tasks include frame detection, character and object identification, and in-depth analysis. Phase 1 is further subdivided into two key subsections which would explain visual perception in depth:

Pixel Scene: Pixel Scene represents a widely used solution strategy for image recognition that is OpenCV. Served as the cornerstone for image preprocessing and manipulation. This powerful and versatile computer vision library offers a broad spectrum of capabilities, enabled to transform raw input images into a refined, noise-free, and optimized format. Here, mentioned below is the deeper information into how the calculation scheme of OpenCV contributes to this phase:

Image Input and Preprocessing: OpenCV efficiently acquires images from users, establishing the foundation for comprehensive preprocessing operations. This process is critical for optimizing the images for subsequent analysis tailored to research's specific requirements.

Noise Reduction: A primary goal in preprocessing phase is noise reduction. OpenCV's toolkit, including techniques such as blurring, smoothing, and denoising, effectively removes noise. This step ensures that the final image is pristine and devoid of artifacts that could interfere with the subsequent analysis, aligning with research's objectives.

Resizing and Standardization: OpenCV plays a vital role in image resizing and standardization, enabling to tailor image dimensions to meet the specific requirements of research. This standardization is essential for achieving consistency throughout the later phases of the analysis.

Color Corrections and Enhancements: In research, OpenCV's capacity for color correction and enhancement was invaluable. Adjustments in brightness, contrast, and color balance were applied to optimize the visual quality of images for downstream tasks.

Single-Glance Detection: The proceeding stage in the Mythological Scene Teller algorithm involves character and object recognition within the pre-processed image. Single-Glance Detection could identify any object present in a mythological scene in just one view. This would make wonder how an algorithm can process an image within a single step. YOLO (You Only Look Once), an object detection algorithm founded on deep neural networks. It offers the ability to perform real-time object detection in images by dividing the input image into a grid and predicting bounding boxes and class probabilities for each grid cell. These bounding boxes provide with essential information, including the class of the object, *X* and *Y* coordinates, and the dimensions of the bounding box (width and height).

To ensure precise detection and localization of the various characters and objects within the image, the custom dataset was meticulously prepared using Roboflow. This custom dataset comprises 1792 images across 18 distinct classes, encompassing 5 scenic classes and 13-character classes. The scenic classes serve to identify the scenes and provide contextual information for subsequent phases. For instance, in a scene depicting Karna's death, the presence of a stuck chariot serves as an iconic symbol and can be classified as a scenic class.

To tailor YOLO to specific requirements, fine-tuned its parameters, defining deeper layers for the MST character recognition model. This model was meticulously trained on the custom dataset, which allowed to precisely recognize characters and extract comprehensive knowledge from the input image. Through the implementation of YOLO, successfully achieved the objective of character recognition and laid the foundation for subsequent phases of the MST algorithm.

5.2 Relationship Establishment

Relationship Establishment focuses on establishing meaningful relationships between the identified characters within the mythological scene, aiming to gain a deeper understanding of the interactions and contextual cues present. The approach involved applying constraints and rules to analyze the interactions and context of the characters, leading to the generation of keywords describing these relationships. By identifying and describing these relationships, the research aimed to provide a more comprehensive interpretation of the mythological scene. This phase incorporated two key algorithms, each discussed in the subsequent subsections.

MRE (Mythological Relationship Establisher): The MRE algorithm, which stands for Mythological Relationship Establisher, formed the cornerstone of the relationship establishment process. In the context of the research, MRE introduced specific approaches to achieve this goal:

Graph-Based Representation: MRE harnessed a graph as its foundational structure, where characters were depicted as nodes, and the connections or edges between them symbolized the relationships within the mythological scene. This graph-based approach allowed for the visual representation of intricate character connections and facilitated the identification of transient relationships.

Dictionary-Based Knowledge: To support the relationship establishment process, MRE maintained a dictionary that stored essential character information, including their identities and relationships. This knowledge base was crucial in guiding MRE during the analysis of textual data within the research.

Named Entity Recognition (NER): NER was a pivotal component of MRE, enabling the recognition and extraction of character entities from the textual paragraphs. NER categorized these entities, such as character names, based on their context within the text. This process involved tokenization, which divided text into individual tokens or words, and Part-of-Speech (POS) tagging to assign labels to these tokens. NER classified entities and identified relationships based on the textual context, enriching the research's understanding of character interactions and relationships.

Contextual Language Processor In addition to MRE, explored an alternative approach for relationship establishment using NLTK, the Natural Language Toolkit. This approach diverged from graph-based representation and involved data extraction and storage in a dictionary without the need for graph structures. In the NLTK method, focused on storing relationships along with the characters in a dictionary. This dictionary allowed to access relationship data by simply referencing the key corresponding to a specific character. This approach offered a more straightforward method to establish character relationships without graph modeling

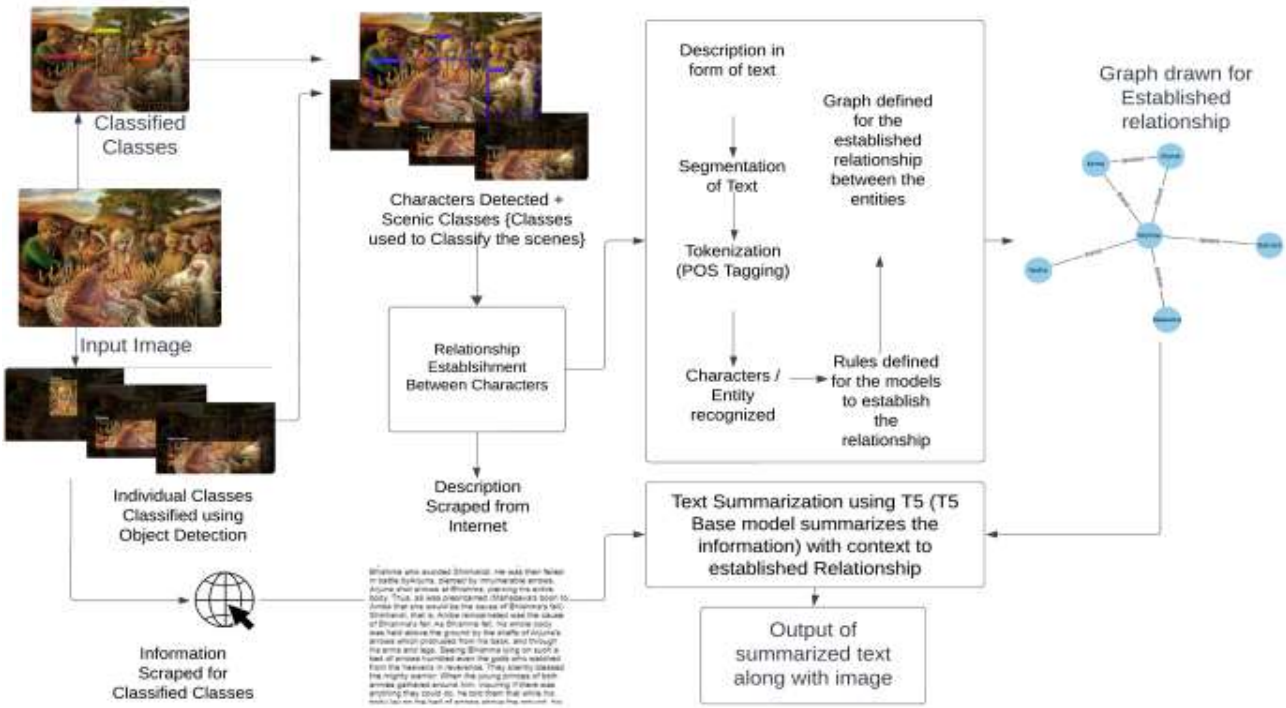
5.3 Structured Data Retrieval

Building upon the keywords and identified relationships generated in the previous phase, there is a need to search relevant information in order to deliver a short and crisp summary. Structured Data Retrieval deep dives into domains where we can extract relevant data with a little bit of noise in it. In Mythological Scene Teller algorithm, to gather relevant information, the use of Web scraping allowed to access a diverse array of information and perspectives related to the mythological scene, further enriching the understanding and providing comprehensive context for subsequent analysis.

Web Harvesting: To execute this phase effectively, employed specialized web scraping tools, with Beautiful Soup taking a prominent role. Beautiful Soup is a Python library that excels in extracting data from websites, articles, scholarly repositories, and other online sources. Its versatile capabilities enabled to navigate websites, identify relevant content, and

extract data in a structured and organized manner.

Data Enrichment and Contextual Understanding: Web scraping techniques played a pivotal role in enriching our understanding of the mythological scene. It allowed to gather a wide range of information, including historical context, cultural interpretations, and scholarly insights. This diverse set of data sources contributed to the research's ability to provide a comprehensive context for further analysis.



5.4 Digestion Algorithm

In the final phase of the research, as the name suggests the digestion algorithm takes input a series of paragraphs of relevant data just like food and meals. The series of lines in paragraph is then processed and shorten to extract most valuable and related information just after the churning process in a stomach. The leveraged advanced text summarization techniques are made use to distill the information acquired through web scraping into concise and enlightening summaries, thereby achieving the primary research objective, offering a comprehensive interpretation of the mythological scene.

TTM (Text Transformer Module): In Mythological Scene Teller algorithm, the Text Transformer Module describes the use of a transformer that is stacking of set of encoders over decoders, to summarize the generated data as information in short. T5 (Text-to-text Transfer Transformer), a pioneering Natural Language Processing (NLP) model, played a pivotal role in the research's success.

Tailored T5 to meet the specific needs of the mythological scene recognition research. T5's unique approach treating all NLP tasks as text-to-text tasks was instrumental, as it allowed to handle various NLP comprehension and generation tasks within a unified framework. T5 works by ingesting text as both input and output. For the research, configured it to take the extensive information extracted from web sources as input and generate concise and coherent summaries as output. With

T5, crafted summaries that were both clear and coherent. These summaries distilled the essence of the mythological scene from the extensive data acquired during web scraping, ensuring the research's success in effectively communicating the narrative's richness.

DFT (Dual-Flow Transformer): In parallel with use of T5, also integrated BERT (Bidirectional Encoder Representations from Transformers), a renowned text summarization model, into the study. BERT's proficiency in comprehending and summarizing textual data effectively was instrumental in enhancing interpretation of the mythological scene. BERT operates by analyzing text bidirectionally, meaning it examines words within a sentence by considering the context of the surrounding words. For the study, this allowed to capture the essence of the mythological scenes by considering the context of the information extracted from web sources. BERT empowered to generate summaries that were not only concise but also coherent, effectively encapsulating the essence of the mythological scene. The information obtained through web scraping was transformed into succinct and engaging summaries that resonated with the research's overarching objectives. (Figure 3) shows the whole overview of the MST algorithm.

5. PERFORMANCE EVALUATION

In the pursuit of mythological scene recognition, MST underwent a comprehensive evaluation. The analysis encompassed three key components: character detection using two distinct datasets, MSTDataset-1 and MSTDataset-2, and text summarization. Each component played a vital role in enhancing the research's interpretive capabilities, culminating in a deeper understanding of mythological scenes. (Table 1) shows the models used and their respective mAP score and precision.

6.1 Character Detection - MSTDataset-1

MSTDataset-1, a rich and diverse dataset comprising 63 classes, posed the first benchmark for character detection models. Three models were employed, each undergoing 300 epochs of training:

YOLOv5: This model achieved a mAP of 30% and a precision rate of 53%, proving effective in character detection. (Figure 4) shows the precision score of YOLO model trained on MSTDataset-1.

FastRCNN (VGG-16): Utilizing the VGG-16 architecture, this model achieved a mAP of 8% and a precision of 36%, offering an alternative approach.

Custom CNN: Tailored for the task, this Custom CNN with 10 convolutional layers and 3 pooling layers achieved a 7% mAP and a remarkable precision rate of 60%.

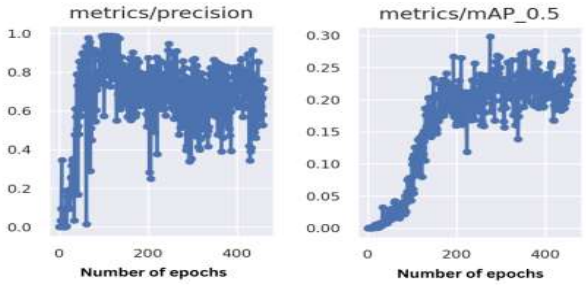


Figure 4. YOLOv5 model trained on MSTDateset-1

Table 1. CNN models for character detection

CNN Models	MST Dataset 1		MST Dataset 2	
	mAP	Precision	mAP	Precision
YOLOv5	30%	53%	62%	70%
FastRCNN (VGG-16)	8%	36%	39%	53%
Custom CNN	7%	60%	48%	60%

6.2 Character Detection - MSTDataset-2

MSTDataset-2, with 18 classes, provided the next testing ground for character detection models. The same three models were evaluated:

YOLOv5: Remarkably, YOLOv5 achieved a mAP of 62% and a high precision rate of 70%, demonstrating its effectiveness. (Figure 5) shows the precision score of YOLO model trained on MSTDataset2.

FastRCNN (VGG-16): For this dataset, the FastRCNN model achieved a mAP of 39% and a precision of 53%, providing an alternative approach.

Custom CNN: Tailored to the dataset, this Custom CNN maintained its effectiveness with a 48% mAP and a precision rate of 60%.

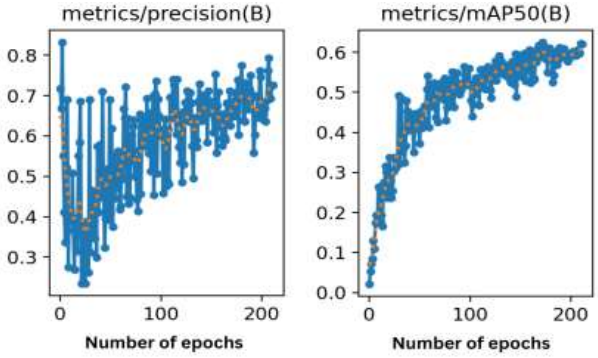


Figure 5. YOLOv5 model trained on MSTDateset-2

6.3 Text Summarization

T5 Base: Generating summaries for 7 out of 10 inputs, T5 Base often introduced unique sentences not present in the original description, enhancing the interpretive capacity of MST.

T5 Small: Although it generated suitable summarizations for 3 out of 10 inputs, some results included less relevant information alongside the summarization.

GPT-2: In 6 out of 10 inputs, GPT-2 successfully provided appropriate summarization, but in some cases, it consisted of direct text extracted from the paragraph rather than a concise summary. (Figure 6) shows the comparison of scores and accuracy of different text summarizing models.

The rigorous evaluation of character detection models using MSTDataset-1 and MSTDataset-2 demonstrated the capabilities of YOLOv5, FastRCNN, and Custom CNN in varying scenarios. The choice of model should be tailored to the dataset and specific requirements. On the text summarization front, T5 Base excelled in providing insightful summaries with added value, while T5 Small and GPT-2 displayed different strengths and limitations.

The initial model of MST-1, based on dataset 1 containing 63 classes and 780 images, achieved a 30% mAP score and 53% for character recognition. However, the information extracted was comparatively vague due to it sourcing information from unauthorized sources.

Subsequently, in an effort to improve the results, the dataset was enhanced, and custom parameters were set for the MST-2 model. The second dataset consisted of 18 classes and 1796 images, resulting in an increased mAP score of 62% and a precision of 70% for character recognition. By utilizing authorized websites, the research was able to scrape data more accurately for various scenes.

In the domain of text summarization, the T5 base BERT summarizer model achieved a precision score of 70%, providing appropriate summaries for the detected classes from the mythological image.

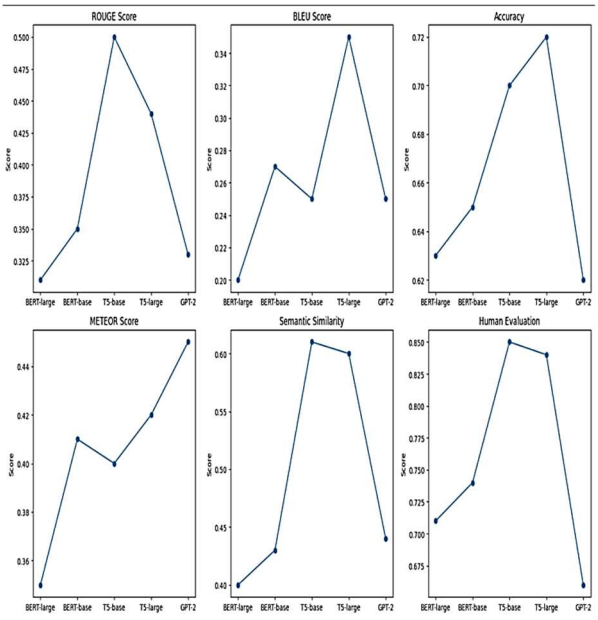


Figure 6. Comparison of scores and accuracy of different text summarizing models

6. CONCLUSION

This research provides insights to successfully implementing a Mythological Scene Recognition for various Indian mythological scenes. In research mentioned a way to extract the authorized description from scenes and the importance of relationship between the characters to get the context for recognizing the scene. The developed application will help to preserve the Mythological history and an innovative way for other people to get knowledge about the historical events providing a real time application not only for tourism, museums but it can also be used in education sector.

References:

A. -A. Liu, Y. Wang, N. Xu, W. Nie, J. Nie and Y. Zhang, ”Adaptively Clustering-Driven Learning for Visual Relationship Detection,” in IEEE Transactions on Multimedia, vol. 23, pp. 4515-4525, 2021, doi: 10.1109/TMM.2020.3043084

A. R. Bharadwaj, S. S. Chandra, D. S. Nair, A. R. Hatim and A. Ravikumar, ”Automated mythological scene recognition using machine learning and graphs,” 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), Amaravati, India, 2020, pp. 1-5, doi: 10.1109/AISP48273.2020.9073474

B. Labinghisa and D. M. Lee, ”A Deep Learning based Scene Recognition Algorithm for Indoor Localization,” 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Korea (South), 2021, pp. 167-170, doi: 10.1109/ICAIIC51459.2021.9415278

B. Oh and J. Lee, ”A case study on scene recognition using an ensemble convolution neural network,” 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 351-353, doi: 10.23919/ICACT.2018.8323752

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, ”Places: A 10 Million Image Database for Scene Recognition,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452-1464, 1 June 2018, doi: 10.1109/TPAMI.2017.2723009

C. Liu, Y. Tao, J. Liang, K. Li and Y. Chen, ”Object Detection Based on YOLO Network,” 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2018, pp. 799-803, doi: 10.1109/ITOEC.2018.8740604

C. R. Dhivyaa, K. Nithya, T. Janani, K. S. Kumar and N. Prashanth, ”Transliteration based Generative Pre-trained Transformer 2 Model for Tamil Text Summarization,” 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-6, doi: 10.1109/ICCCI54379.2022.9740991

H. Seong, J. Hyun, H. Chang, S. Lee, S. Woo and E. Kim, ”Scene Recognition via Object-to-Scene Class Conversion: End-to-End Training,” 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-6, doi: 10.1109/IJCNN.2019.8852040

H. Shalma and P. Selvaraj, ”Deep-Learning Based Object Detection and Shape Recognition in Multiple Occluded Images,” 2022 International Conference on Data Science, Agents Artificial Intelligence (ICDSAAI), Chennai, India, 2022, pp. 1-7, doi: 10.1109/ICDSAAI55433.2022.10028848

H. Zeng, X. Song, G. Chen and S. Jiang, ”Learning Scene Attribute for Scene Recognition,” in IEEE Transactions on Multimedia, vol. 22, no. 6, pp. 1519-1530, June 2020, doi: 10.1109/TMM.2019.2944241

J. Tao, Y. Gu, J. Sun, Y. Bie and H. Wang, ”Research on vgg16 convolutional neural network feature classification algorithm based on Transfer Learning,” 2021 2nd China International SAR Symposium (CISS), Shanghai, China, 2021, pp. 1-3, doi: 10.23919/CISS51089.2021.9652277

J. Wu et al., ”Progressive Guided Fusion Network With Multi-Modal and Multi-Scale Attention for RGB-D Salient Object Detection,” in IEEE Access, vol. 9, pp. 150608-150622, 2021, doi: 10.1109/ACCESS.2021.3126338

J. Zhu and H. Wang, ”Multiscale Conditional Relationship Graph Network for Referring Relationships in Images,” in IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 752-760, June 2022, doi: 10.1109/TCDS.2021.3079278

K. Aigo, T. Tsunakawa, M. Nishida and M. Nishimura, ”Question Generation using Knowledge Graphs with the T5 Language Model and Masked Self-Attention,” 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, 2021, pp. 85-87, doi: 10.1109/GCCE53005.2021.9621874

Li, Y. Li, Y. Li, M. Li and X. Xu, ”YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection,” in IEEE Access, vol. 9, pp. 141861-141875, 2021, doi: 10.1109/ACCESS.2021.3120870

L. Liu, Y. Wang and W. Chi, ”Image Recognition Technology Based on Machine Learning,” in IEEE Access, doi: 10.1109/ACCESS.2020.3021590

Lunn, Stephanie Zhu, Jia Ross, Monique. (2020). Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice. 10.1109/FIE44824.2020.9274270

M. Han, S. Li, X. Wan and G. Liu, ”Scene Recognition with Convolutional Residual Features via Deep Forest,” 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, China, 2018, pp. 178-182, doi: 10.1109/ICIVC.2018.8492736

M. R. Suryakusuma, M. Faqih Ash Shiddiq, H. Lucky and I. A. Iswanto, ”Investigating T5 Generation Neural Machine Translation Performance on English to German,” 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), Jakarta Selatan, Indonesia, 2023, pp. 12-15, doi: 10.1109/ICIMCIS60089.2023.10349061

M. Yavartanoo, S. -H. Hung, R. Neshatavar, Y. Zhang and K. M. Lee, ”PolyNet: Polynomial Neural Network for 3D Shape Recognition with PolyShape Representation,” 2021 International Conference on 3D Vision (3DV), London, United Kingdom, 2021, pp. 1014-1023, doi: 10.1109/3DV53792.2021.00109

N. Darapaneni, R. Prajeesh, P. Dutta, V. K. Pillai, A. Karak and A. R. Paduri, ”Abstractive Text Summarization Using BERT and GPT-2 Models,” 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommuncation (IConSCEPT), Karaikal, India, 2023, pp. 1-6, doi: 10.1109/IConSCEPT57958.2023.10170093

P. C. Blaud, P. Chevrel, F. Claveau, P. Haurant and A. Mouraud, ”ResNet and PolyNet Based Identification and (MPC) Control of Dynamical Systems: A Promising Way,” in IEEE Access, vol. 11, pp. 20657-20672, 2023, doi: 10.1109/ACCESS.2022.3196920

Q. Bi, K. Qin, H. Zhang and G. -S. Xia, ”Local Semantic Enhanced ConvNet for Aerial Scene Recognition,” in IEEE Transactions on Image Processing, vol. 30, pp. 6498-6511, 2021, doi: 10.1109/TIP.2021.3092816

T. Zheng et al., ”Scene Recognition Model in Underground Mines Based on CNN-LSTM and Spatial-Temporal Attention Mechanism,” 2020 International Symposium on Computer, Consumer and Control (IS3C), Taichung City, Taiwan, 2020, pp. 513-516, doi: 10.1109/IS3C50286.2020.00139

W. Fang, L. Wang and P. Ren, ”Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments,” in IEEE Access, vol. 8, pp. 1935-1944, 2020, doi: 10.1109/ACCESS.2019.2961959

X. Wu, D. Hong and J. Chanussot, ”Convolutional Neural Networks for Multimodal Remote Sensing Data Classification,” in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-10, 2022, Art no. 5517010, doi: 10.1109/TGRS.2021.3124913

Yuan, Xiaohui Qiao, Zhinan Meyarian, Abolfazl. (2021). Scale Attentive Network for Scene Recognition. Neurocomputing. 492. 10.1016/j.neucom.2021.12.053

Z. Qu, G. Xiao, N. Xu, Z. Diao and H. Jia-Zhou, ”A novel night vision image color fusion method based on scene recognition,” 2016 19th International Conference on Information Fusion (FUSION), Heidelberg, Germany, 2016, pp. 1236-1243

Z. Wang and Z. Li, ”Person Sensor-Aided Scene Recognition and Understanding Based on CG Technology,” 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 60-63, doi: 10.1109/ICICT48043.2020.9112445