IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104082 |

Hierarchical Text-Conditional Image Generation with CLIP Latents

Mr. S. Y. Divekar [*1], Mr. P. P. Soma[*2], Ms. T. D. Kalapure[*3] , Mr. C. P. Tambe[*4], Ms. D. R. Pukale[*5]

[*1]Lecturer, Computer Engineering Department, AISSMS Polytechnic, Pune, Maharastra, India

[*2]Student, Computer Engineering Department, AISSMS Polytechnic, Pune, Maharastra, India

[*3]Student, Computer Engineering Department, AISSMS Polytechnic, Pune, Maharastra, India

[*4]Student, Computer Engineering Department, AISSMS Polytechnic, Pune, Maharastra, India

[*5]Student, Computer Engineering Department, AISSMS Polytechnic, Pune, Maharastra, India

ABSTRACT: In various real-time applications, several assisted services are provided by human-robot interaction (HRI). The concept of convergence of a three-dimensional (3D) image into a plane-based projection is used for object identification via digital visualization in robotic systems. Recognition errors occur as the projections in various planes are misidentified during the convergence process. These misidentifications in the recognition of objects can be reduced by an input processing scheme dependent on the projection technique. The conjoining indices are identified by projecting the input image in all possible dimensions and visualizing it. A machine learning algorithm is used for improving the processing speed and accuracy of recognition. A labeled analysis is used for the segregation of the intersection without conjoined indices. Errors are prevented by identifying the noncorrelating indices in the projections of possible dimensions. The inputs are correlated with related inputs that are stored with labels thereby preventing the matching of the indices and deviations in the planes. Error, complexity, time, and recognition ratio metrics are verified for the proposed model.

KEYWORDS: human-robot interaction, Virtual reality modeling language (VRML).

I. INTRODUCTION

To access any object without human involvement within the specified period, human-robot interactions are used. Virtual reality objects are defined using the three-dimensional (3-D) image sequences in the Virtual reality modeling language (VRML) [1]. Virtual objects and their scenes may be approached using a VRML programming scheme. Based on the object, a virtual robot is generated for supporting the workspace environment. In web settings, a series of virtual images may be generated by the user with the help of VRML which offers connectivity with a seemingly 3D scene by spinning, turning, watching, or otherwise [2]. Currently, multipurpose environments or virtual reality headsets are used by conventional augmented reality devices for the creation of compelling emotions, sounds, or images by stimulating the actual appearance of a person in a virtual world. Communication between the user and the virtual functions and elements in the simulated environment is made possible using augmented reality devices. From the sensed object, quantitative data collection may be derived and the location of the object may be observed in this area [3]. The object is acquired from the surrounding in a 3D view for the visualization process. Conversion processes are implemented as the object in a 3D view may not be recognized by the robot. Movement and position of the object in space may be detected using a robot configuration space. The changes in position are processed and determined every time by sequential monitoring of objects [4]. Concerning the previous history, the nature of the object is detected by the sensor and the object is visualized. The result of the object in space is derived by implementing matching schemes for finding the object.

II. METHODOLOGY

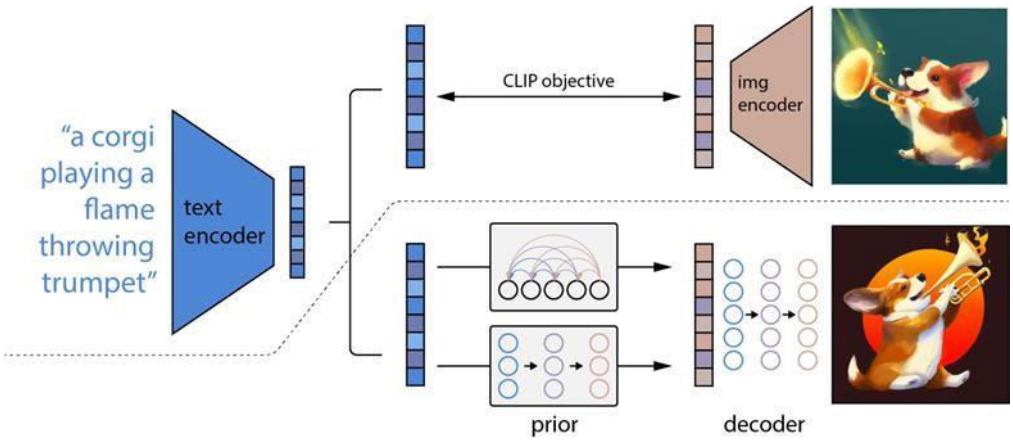
THE OBJECTS IDENTIFIED FROM THE SURROUNDINGS ARE SHARED BETWEEN HUMANS AND ROBOTS UTILIZING HUMAN-ROBOT INTERACTION. 3D-BASED IMAGES ARE MORE COMMONLY OBTAINED FROM THE ENVIRONMENT. HOWEVER, THE

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104082 |

DETECTION OF THESE IMAGES IS DIFFICULT FOR ROBOTS. FOR THIS PURPOSE, WE DEPLOY A 3D TO 2D CONVERSION PROCESS FOR ACCURATE IDENTIFICATION OF THE OBJECT AND TO PROVIDE RESULTS TO HUMANS. FIGURE 1 REPRESENTS THE ARCHITECTURE OF THE



. **Figure 1:** architecture of the proposed HRI model.

Machine learning algorithms and projection-dependent input processing features are used for image recognition using specific entities and prevention of incorrect identification. Based on joint indexes, recurrent analysis is performed for the deletion of detected components. Timeline recurrent analysis is used for describing and denoting the identification based on the prior analysis. The direction of the object is acquired by the robot from a specific region from which the sense of the object is used for validation of the input. The plane calculation is used for the detection of the position and indices of the object. These factors are detected based on the 3D to 2D object that is obtained. Two types of optimization is performed for the indices. The first type of optimization is on the principle that multiple positions may be available for objects that reside on a plane. The second optimization technique is based on the principle that the object relies on a different plane if the object's position does not vary. Optimization of indices is performed for the detection and calculation of these constraints. The plane in which the object resides and its position are the two major constraints based on which the object is selected from the plane and output is produced.

III. MODELING AND ANALYSIS

Our training dataset consists of pairs (x, y) of images x and their corresponding captions y. Given an image x, let zi and zt be its CLIP image and text embeddings, respectively.

We design our generative stack to produce images from captions using two components:

A prior P(zi |y) that produces CLIP image embeddings zi conditioned on captions y.

A decoder P(x|zi , y) that produces images x conditioned on CLIP image embeddings zi (and optionally text captions y).

The decoder allows us to invert images given their CLIP image embeddings, while the prior allows us to learn a generative model of the image embeddings themselves. Stacking these two components yields a generative model P(x|y) of images x given captions.

$$\mathbf{P(x|y) = P(x, zi\ |y) = P(x|zi\ , y)P(zi\ |y).}$$

The first equality holds because zi is a deterministic function of x. The second equality holds because of the chain rule. Thus, we can sample from the true conditional distribution P(x|y) by first sampling zi using the 3 prior, and then sampling x using the decoder. In the following sections, we describe our decoder and prior stacks.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104082 |

IV. RESULTS AND DISCUSSION

3D images from a source are used for the analysis of the performance of the proposed machine learning algorithm-based input processing model. 1000 sample images from various classes are considered 3D models. Verification of correlation and matching is performed using a testing and training dataset with 500x2 images. MATLAB software is used for analyzing these images. For every matching instance, 50 images are trained. Input correlation and verification of indices are performed for every instance. Error, processing complexity, processing time, and recognition ratio are analyzed for estimating the efficiency of the proposed method. In all tests, the association factor and several items are verified. Existing schemes like a multi-scale convolutional neural network, Context-Assisted 3D, Conventional voxel-based occupancy grid, radio-frequency identification, support vector regression, and Dynamic Statistical Parametric Mapping are compared.

V. CONCLUSION

In human-robot interaction settings, 3D objects are identified using the proposed projection-based input processing model. Based on layer selection and vector representation, a 3D object is converted into various planes of 2D. Two vectors are defined and their cross-product is determined based on the element-wise vector product in a 3D space. The external product with an n-dimensional abstract vector product, the name vector is combined with wedge notation. With respect to the concurrent and prior detection process, the labeling of indices is performed during the conversion process. The errors are reduced and identification and mitigation of non-correlating points from different planes are also done based on the indices matching instances. Detection of a region of interest in the indices data projection helps in identifying the variations in the plane. Periodic matching updates are performed along with the correlation and matching process concurrently. The error, complexity, and processing time are reduced while improving the recognition ratio using the proposed technique based on experimental analysis. Future work is directed toward improving these values further and enhancing the performance of the model during real-time implementation.

REFERENCES

[1] Luo, R. C., & Wu, X. (2014, March). Real-time gender recognition based on 3d human body shape for humanrobot interaction. In Proceedings of the 2014 ACM/IEEE international conference on Humanrobot interaction (pp. 236-237).

[2] Waldherr, S., Romero, R., & Thrun, S. (2000). A gesture based interface for human-robot interaction. Autonomous Robots, 9(2), 151-173.

[3] Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., ... & Mao, J. (2017). A facial expression emotion recognition based human-robot interaction system

[4] Li, X. (2020). Human–robot interaction based on gesture and movement recognition. Signal Processing: Image Communication, 81, 115686.

[5] Mazhar, O., Ramdani, S., Navarro, B., Passama, R., & Cherubini, A. (2018, October). Towards real-time physical human-robot interaction using skeleton information and hand gestures. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1-6). IEEE.

[6] Li, J., Mi, Y., Li, G., & Ju, Z. (2019). Cnn-based facial expression recognition from annotated rgb-d images for human–robot interaction. International Journal of Humanoid Robotics, 16(04), 1941002.

[7] Filippini, C., Perpetuini, D., Cardone, D., Chiarelli, A. M., & Merla, A. (2020). Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: a review. Applied Sciences, 10(8), 2924.

[8] Chen, L., Zhou, M., Su, W., Wu, M., She, J., & Hirota, K. (2018). Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. Information Sciences, 428, 49-61.

[9] Du, G., Chen, M., Liu, C., Zhang, B., & Zhang, P. (2018). Online robot teaching with natural human– robot interaction. IEEE Transactions on Industrial Electronics, 65(12), 9571-9581.

[10] Deng, J., Pang, G., Zhang, Z., Pang, Z., Yang, H., & Yang, G. (2019). cGAN based facial expression recognition for human-robot interaction. IEEE Access, 7, 9848-9859.

[11] Fang, B., Sun, F., Liu, H., & Liu, C. (2018). 3D human gesture capturing and recognition by the IMMUbased data glove. Neurocomputing, 277, 198-207.

[12] Shridhar, M., & Hsu, D. (2018). Interactive visual grounding of referring expressions for human-robot interaction. arXiv preprint arXiv:1806.03831.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104082 |

[13] Shakya, S. (2019). Virtual restoration of damaged archeological artifacts obtained from expeditions using 3D visualization. Journal of Innovative Image Processing (JIIP), 1(02), 102-110.

[14] Dhaya, R. (2020). Improved Image Processing Techniques for User Immersion Problem Alleviation in Virtual Reality Environments. Journal of Innovative Image Processing (JIIP), 2(02), 77-84.

[15] Ranganathan, G. (2020). Real Life Human Movement Realization in Multimodal Group Communication Using Depth Map Information and Machine Learning. Journal of Innovative Image Processing (JIIP), 2(02),

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details