

The Feasibility of Network Data Transmission Reducing based on Vision Language Model (VLM)

Ming-Shen Jian, Ming-Hsun Tsai, Chun-Ho Cheng, Jay-Hong Cheng, Bo-Yuan Huang, Shih-Fang Huang, Cai-Jhen Li

CCIS Lab., Dept. of CSIE, National Formosa University, Yunlin County, Taiwan 632

jianms@nfu.edu.tw, sean408031015@gmail.com, neil990006@gmail.com, 41143156@gm.nfu.edu.tw,
41043225@gm.nfu.edu.tw, 41243205@gm.nfu.edu.tw, 41243102@gm.nfu.edu.tw

Abstract— This study verifies that transmitting plain text descriptions, instead of image files, can effectively reduce network transmission when conveying moods or concepts through images. The proposed approach requires only 1–1.25% of the data size of the original image. A data structure integrating a basic vision-language model with relational links between multiple objects is implemented. Experimental results confirm the feasibility of translating images into descriptive text for network transmission. Both reduced data load and preserved semantic content for efficient communication would be available.

Keywords— Image Recognition, Artificial Intelligence, Vision Language Model, Network

I. INTRODUCTION

Virtual world technology has gained significant attention in recent years due to its wide range of applications [1,2]. To improve the performance of automation systems, machine learning and artificial intelligence (AI) are increasingly integrated into digital twin frameworks, enhancing the accuracy of recognition, identification, and process optimization. With the use of hardware devices such as virtual reality (VR) headsets or smart glasses, users can interact with virtual objects and communicate with others in a shared virtual environment via the Internet. Furthermore, the application of augmented reality (AR) enables interaction with physical objects within the virtual world [3,4]. By utilizing camera-based intelligent modules, physical objects can be detected, recognized, and converted into textual data for subsequent operations or applications [5]. Through the recognition of objects, operators, gestures, and even spatial positions, eXtended reality (XR) technologies can seamlessly merge physical objects with virtual environments, creating an integrated and immersive visual experience.

With the rapid advancement of artificial intelligence (AI) technologies, applications capable of generating images or melodies from textual descriptions have become increasingly prevalent [5–7]. In such services, users provide descriptive text about objects, scenes, or emotions. Then, the AI system generates diverse outputs such as images or musical pieces based on the trained models. These models are developed by using large-scale datasets compiled from multiple sources including user contributions and existing databases. Consequently, a single description can create multiple plausible and contextually varied results which reflect the diversity inherent in the training data.

On the other hand, object recognition often involves translating visual information into textual descriptions. Factors such as recognition accuracy, object segmentation, and the relative importance of objects within the image can significantly influence the resulting description. It will can lead to various interpretations. As a result, the same image may yield different explanations depending on how these elements are prioritized. To generate descriptions that align more closely with human interpretations would become an important issue.

In addition to the object recognition by AI, to transmit the data from one terminal to another through network would also cost time and bandwidth. Generally speaking, an image as a jpeg file would be several kilo bytes (KB) with low quality or mega bytes (MB) with high quality. However, the text data for transmission would be only few bytes (B). It means that the time latency of data transmission through network could be reduced by using text data. Therefore, based on the suitable Vision Language Model for image and object

recognition and relationship defining, the total amount of data transmitted can be decreased.

The main contributions of this research can be shown as follows:

1. By integrating artificial intelligence and computer vision, the proposed system achieves body and gesture recognition of the operator without the need for handheld controllers or wearable devices. This design allows users to interact with the system in a more natural and intuitive manner. It significantly improves ease of operation and user experience.
2. Data can be transmitted through the network via text data packet. The latency of the data transmission can be decreased. In addition, the recognition result and the description can be more easily reused by any other applications or program. Users in different locations can collaboratively operate or share the concept of the vision easily.
3. The proposed system considers the Mandarin language which is more difficult in grammar for the description corresponding to the image recognition. Any further translate between different languages can be done.

The organization of this research is as follows: the research related works are presented in section 2. Section 3 is the proposed system. Section 4 shows verification. Conclusion is given in section 5.

II. RELATED WORKS

Interaction with virtual objects in a virtual environment traditionally relies on IoT devices or sensors to detect physical states, such as human motion, acceleration (G-sensor) data, or inputs from joystick controllers. Recently, cameras become a common tool in AI-based computer vision applications [2]. By detecting and recognizing user actions or gestures, AI modules can determine and execute appropriate responses. In addition, according to the training data, the AI application could find the position of the recognized objects. Based on the position, location, etc., the relationship between recognized objects can be defined [5]. Then, the recognized data can be transmitted to remote servers for additional processing. Applications built on virtual environments, including virtual reality

(VR), augmented reality (AR), and extended reality (XR), can then utilize this processed data to deliver corresponding operations or interactive responses [1-4].

Generating an object that corresponds to a textual description requires prior definition of the object's features. Traditionally, object recognition has relied on extensive training with large-scale image datasets. By labeling diverse objects within such datasets, it becomes possible to achieve accurate object recognition and segmentation [9]. According to the object recognition results, the relative position and the size of the segmentation corresponding to each recognized object included in the current image can be marked and recorded.

To identify and recognize the objects in the images or streaming, many applications and systems are developed based on YOLO (You Only Look Once) algorithm [10]. Applications for music instrument [11], underwater objects [12], and target object with less different in color [13] can be recognized via pre-trained data or functions.

Since the objects, even the gesture or posture, are recognized, how to translate these nouns or states into the humanized language description is an important issue. Vision language model recently is mentioned for giving the description corresponding to the caught images or visions [5,14]. Via suitable theories, the verb or adjective connections between different objects can be defined. Object relations such as subject and possessive can also be defined. Therefore, the vision or image could be translated into sentence or text.

III. PROPOSED SYSTEM

In this research, the proposed system could be divided into different partitions: Object Segmentation and Recognition, Vision Language Model based object relation construction.

Object Segmentation and Recognition

In this research, we use YOLO and additional training data related to the pre-labelled object. The trained YOLO model could be applied to embedded devices (such as PC, Raspberry, Jetson Nano, etc.) and utilize e resources (CPU and GPU) to achieve low-latency object detection and recognition. In this research, we also use an object-oriented-similar

database to store information and data for artificial intelligence correspond to the information about the recognized object.

When an object is recognized, according to the object-oriented-similar data structure, there are multiple attributes such as object name (noun), color, related verb or adjective, affiliation or position, etc., required and filled based on the recognition results. In other words, considering the possible multiple attributes of various objects, AI model should detect and recognize as much as possible for these required data.

Vision Language Model based object relation construction

Since the available data corresponding to the attributes of each object now is recorded, to construct the relation or connection between different objects is the next step.

In this research, we assume that the human being will be the main subject term of the object-oriented-similar data structure between multiple objects. In other words, if there is any person recognized in the image, the root node of the relation structure will be the recognized individual. Otherwise, the object with the biggest area will be set as the root node of the relation structure. If there is no person or huge object as the main subject term, the object which locates at the centre is set as the root node of the relation structure. It means that there would be multiple subject terms especially when there are multiple persons.

When the subject terms are decided, other recognized objects will be considered. According to the attributes of the main subject term, in addition to the physical properties, the possible addition attributes now will be checked. Assume that the area of each recognized object (main subject term) as 1. Then, by comparing the size or area of other recognized objects, we could calculate and evaluate the object (exclude the main subject term) with the biggest area which overlap on the area of main subject term. According to name and attributes of this biggest recognized object, the connection between main subject term and the current object could be established if the attributes of main subject term can be satisfied. Furthermore, by evaluating the overlap ratio between main subject term and the

current object, the suitable verb or adjective can be used for the connection. Then, the current object could be the leaf node of the root node, main subject term. The steps of the proposed Vision Language Model based object relation construction can be shown as follows:

1. recognize all objects and define the main subject term
2. set the root node and check the corresponding attributes.
3. evaluate the current constructed largest object (excluding the root node object) which has the most overlap area of the root node object.
4. check the attributes of the root node object for the connection conditions. If the condition is satisfied or matched, system would decide the suitable verb or adjective via evaluating the ratio and position of overlap area. Then, the connection between current object (leaf node) and root node object is established. Then, the tree structure is established.
6. if there are more other and smaller objects which have overlap with the leaf node object, we will construct the new tree structure by taking the current leaf node as the new root, the go step 2.
7. if there are more other and smaller objects only overlap with the current root node object, go step 3.
8. after constructing the whole tree structure, the description of the image can be given.

IV. VERIFICATION RESULTS

In this research, in order to recover the text content to the image, we use DeeVid AI and Adobe Firefly. Considering the description for the same image could be various, this research uses OpenAI's CLIP model as a reference for evaluation. To ensure fairness in the evaluation and prevent the use of targeted training to improve accuracy, the CLIP model will not be fine-tuned in the experiment, and only the ViT-B/32 pre-trained model will be used [14].

Scenario 1

In the verification scenario 1, we input the image shown as figure 1 (a). After recognition and construction, we have the sentences from the proposed system and CLIP as "A person wearing pink clothes is holding a black camera" and "A

young woman in a pink hoodie takes a picture”. Based on the two description in English and the original description in Mandarin, we use different AI model to generate image, left side image is generated based on our system according to the English sentence, center image is based on the description from CLIP, and right side is the image generated according the Mandarin sentence. In figure 1(b), the DeeVid AI generate the images according to the both sentences. In figure 1(c), the Adobe Firefly generate the images according to the all these three sentences. The left side image is generated based on our system according to the English sentence, center image is based on the description from CLIP, and right side is the image generated according the Mandarin sentence.

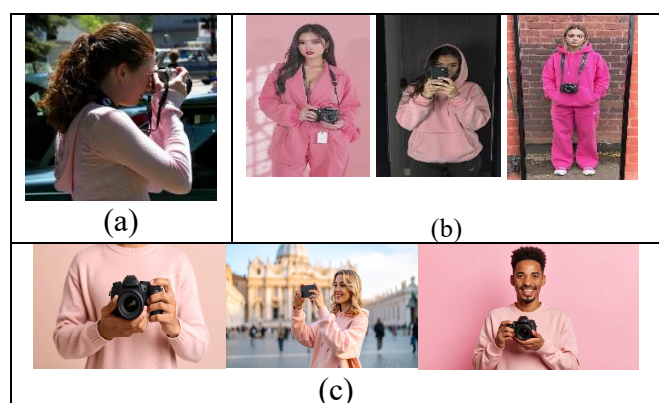


Figure 1. The example of VLM for data translation and recovery.

Scenario 2

In the verification scenario 2, we input the image shown as figure 2 (a). After recognition and construction, we have the sentences from the proposed system and CLIP as “A person wearing blue clothes, a white striped hat, and black glasses is sitting on a brown bench.” and “An old man with a long beard and jean jacket sits on a park bench”. Based on the two description in English and the original description in Mandarin, we use different AI model to generate image. In figure 2(b), the DeeVid AI generate the images according to the three description sentences (ours in English, CLIP, ours in Mandarin). In figure 1(c), the Adobe Firefly generate the images according to the three description sentences (ours in English, CLIP, ours in Mandarin).

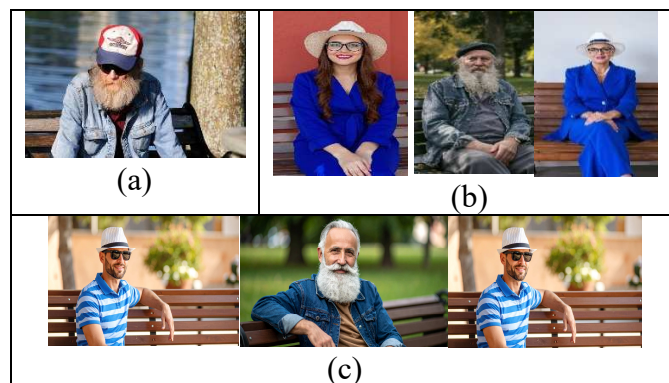


Figure 2. The example of VLM scenario 2 for data translation and recovery.

However, if we increase the attributes of the objects for more information, the text in scenario 2 could be changed as “A man about 70 years old occupies 40% of the picture. He has a white beard, wears a blue denim jacket, his hands are facing down, he wears a white hat and black glasses, and he sits with his back to the bench with brown stripes”. Then, we used both English in figure 3 (b) and Mandarin description in figure 3 (c) for the image generation. Obviously, when more attributes given to AI, to similarity between images would be increased.

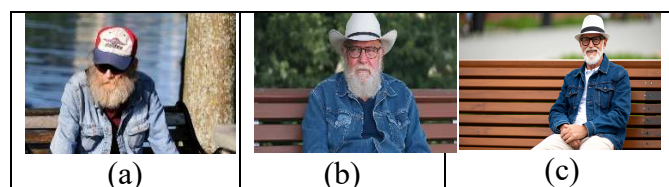


Figure 3. The example of VLM scenario 2 for additional attributes in image description.

By using more attributes, the total size of the images would be 198KB with 16:9 of image (b) and 1.07MB with 16:9 of image (c). Even we compress the original image files, it still is 18.1KB. Instead of images, the description in plain text files would be only 227B in English and 188 in Mandarin. Comparing with the original image, the total transmitted data would be only 1~1.25% of original image.

V. CONCLUSIONS

According to the verification, if we only need to present or express a mood or concept via image, to

transmit plain text instead of image itself could be the possible solution for reducing the network transmission. This research also shows the feasibility of the data structure which deploys the basic vision language model with the relationship connection between multiple objects. Through the translated plain text instead of the image file, the total amount of data transmitted by network can be reduced.

ACKNOWLEDGEMENT

Thank you for the contribution and assistance of Cloud Computing and Intelligent System Lab. in National Formosa University, Taiwan.

REFERENCES

- [1] A. Arsalan, T. Umer, R. A. Rehman, A. M. Abdelmoniem and S. Mumtaz, "Toward Collaborative Intelligence in Digital Twin-based Federated Deep Reinforcement Learning: Applications, Case Studies and Challenges," IEEE Network, 2025, pp.1-1, doi: 10.1109/MNET.2025.3592547.
- [2] Ming-Shen Jian, Chun-Ho Cheng, Jay-Hong Cheng, Ming-Hsun Tsai, Wen-Yu Chung, Chi-Hua Chang, "Multi-Users Extended Reality Remote Collaborative Operation Application based on Digital Twins 3D Projecting," 27th International Conference on Advanced Communications Technology (ICACT), 2025, pp.283-288, doi: 10.23919/ICACT63878.2025.10936796
- [3] M. C. Hsieh and Y. H. Lin, "VR and AR applications in medical practice and education," The Journal of Nursing, 64(6), 12-18, 2017doi:10.6224/JN.000078
- [4] M. -S. Jian, Y. -C. Wang, B. -H. Wu and Y. -E. Cheng, "Hybrid cloud computing for user location-aware augmented reality construction," 20th International Conference on Advanced Communication Technology (ICACT), 2018, pp. 1-1, doi: 10.23919/ICACT.2018.8323691.
- [5] H. -Y. Huang, M. -H. Tsai and M. -S. Jian, "Multiple Merged Structures Based on Image Recognition for Converting Application of Natural Language Artificial Intelligence Service," 26th International Conference on Advanced Communications Technology (ICACT), 2024, pp. 158-162, doi: 10.23919/ICACT60172.2024.10471930.
- [6] N. Capece, G. Manfredi, V. Macellaro and P. Carratù, "An easy Hand Gesture Recognition System for XR-based collaborative purposes," 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE), Rome, Italy, 2022, pp. 121-126, doi: 10.1109/MetroXRINE54828.2022.9967592.
- [7] T. Akiyoshi, Y. Shimizu, Y. Takahama, K. Nagata and T. Sawabe, "Hype D-Live: XR Live Music System to Entertain Passengers for Anxiety Reduction in Autonomous Vehicles," 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, 2023, pp. 148-156, doi: 10.1109/ISMAR59233.2023.00029.
- [8] M. -S. Jian, Y. -C. Wang, B. -H. Wu and Y. -E. Cheng, "Hybrid cloud computing for user location-aware augmented reality construction," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 1-1, doi: 10.23919/ICACT.2018.8323691.
- [9] A. Torralba, B. C. Russell and J. Yuen, "LabelMe: Online Image Annotation and Applications," in Proceedings of the IEEE, vol. 98, no. 8, pp. 1467-1484, Aug. 2010, doi: 10.1109/JPROC.2010.2050290.
- [10] N. Capece, G. Manfredi, V. Macellaro and P. Carratù, "An easy Hand Gesture Recognition System for XR-based collaborative purposes," 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE), Rome, Italy, 2022, pp. 121-126, doi: 10.1109/MetroXRINE54828.2022.9967592.
- [11] T. Akiyoshi, Y. Shimizu, Y. Takahama, K. Nagata and T. Sawabe, "Hype D-Live: XR Live Music System to Entertain Passengers for Anxiety Reduction in Autonomous Vehicles," 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, 2023, pp. 148-156, doi: 10.1109/ISMAR59233.2023.00029.
- [12] U. Žákovská and D. Sedláček, "Studying the Role of Non-Player Users (Spectator and Operator) in XR Single and Multiplayer Experiences," 2023 14th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 2023, pp. 000129-000134, doi: 10.1109/CogInfoCom59411.2023.10397540.
- [13] B. Aresh et al., "Integration of Extended Reality (XR) in Non-Native Undergraduate Programmes," 2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Kuala Lumpur, Malaysia, 2023, pp. 285-292, doi: 10.1109/SKIMA59232.2023.10387333.
- [14] M.-H. Tsai, "Multiple Merged Structures Image Recognition To Natural Language System Based On An Improved YOLO Algorithm," Master thesis, 2025. <https://hdl.handle.net/11296/tx3reb>



Ming-Shen Jian was born in Kaohsiung City, Taiwan in 1978. He received the B.S. from the National Chiao Tung University, HsinChu, and Ph.D degrees in Computer Science and Engineering from the National Sun Yat-sen University, Kaohsiung, Taiwan in 2007. From 2018, he was an Associate Professor and director with the National Formosa University Cloud Computing and Intelligent System Laboratory. Currently he is also an IEEE Senior Member. Since 2009, he has been an Assistant Professor with the Computer Science and Information Engineering Department, National Formosa University. He is the author of four books, more than 50 articles, and at least 15 invention patents. His research interests include IOT development and application, Big Data, Optimal Solution, Intelligent System, and Cloud Computing. He was a Secretary of the Taiwan Association of Cloud Computing. Dr. Jian was a recipient of the IEEE sponsored international conference Paper Award in 2016, 2017, and 2018.



Ming-Hsun Tsai received his master degree in Computer Science and Information Engineering at National Formosa University on 2025. His research interests are in the area related to object recognition and segmentation, and image processing. He was the main researcher about the image progressing in the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023.



Chun-Ho Cheng currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to AI and Docker container services, digital twins, and IoT system. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2022.



Jay-Hong Cheng currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. His current research interests are in the area related to IoT and Docker container services, and multimedia system. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2022.



Bo-Yuan Huang received his currently bachelor degree in Computer Science and Information Engineering at National Formosa University. He currently is the master degree student. His current research interests are in the area related to Cloud Computing and Docker container services, and intelligent system. He joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2021.



Shih-Fang Huang currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. Her current research interests are in the area related to digital twins and multimedia services. She joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023.



Cai-Jhen Li currently is a bachelor degree student of Dept. Computer Science and Information Engineering at National Formosa University. Her current research interests are in the area related to IoT and digital twins services. She joins the Cloud Computing and Intelligent System Lad. (CCIS Lab.) from 2023.