Invisible Pixel: Short Video Narratives from Machine Perspective

Junlin Zhu, Juanjuan Long, Yingjing Duan, and Wenxuan Zhao

Fig. 1. The initial interface of system Invisible Pixel

Abstract— Short video software has become a new farming tool for Chinese farmers to fight against poverty in recent years. The Internet has connected the public cultural space between urban and rural areas, social media has transformed personal expression into public communication, and the public has rediscovered individual narratives from remote areas. *Invisible Pixel* is an interactive web page developed based on the background of poverty alleviation through the Internet in China. The text of short videos has been converted into images through machine learning to create a machine-perspective data experience, in contrast to people's daily viewing of short videos, to explore how computer technology will affect social media in the future.

Index Terms—Poverty alleviation through Internet, Short video, Machine learning, Data narrative

1 Introduction

2020 is the final year of the fight against poverty through joint efforts in China, and the Internet plays a crucial role in it. The 47th China Internet Network Information Center Statistical Report shows that as of December 2020, the size of China's Internet users reached 989 million, and the Internet penetration rate reached 70.4%, including 55.9% in rural areas. The scale of short video users is also imposing, with almost 7 out of 10 Internet users using short video software. According to the official data released by the short video platform "Kwai," as of August 2019, there were more than 5 million users from the national level of poor counties in China, and the total number of videos they released exceeded 1.1 billion, with annual sales reaching 19.3 billion yuan. The data truly reflects the current social situation: China's vast rural areas, especially the poor ones, have gained unprecedented attention resources, enabling them to be truly visible to the outside world and establish positive connections, which is helpful to facilitate rural vitalization; this rural narrative, separated from the consumerist and city-centric, has broken through the class barriers. It rapidly expanded the cultural intimacy between urban and rural areas and satisfied the curiosity of other classes to pry into rural narratives.

With the popularity of short videos, the narrative form breaks the traditional narrative framework of poverty alleviation, and poor groups directly become the narrative subjects. Undoubtedly, computer technology has expanded who and how we watch. So, how will technological innovation further affect social media? If, as Heidegger argues, science places everything opposite to human beings as an object, can the current short video-based individual narratives be objectified with algorithms? In what form will it be presented? Will it be more attractive? With such thoughts and purposes, we try to simulate the data experience of social media from a machine perspective to respond.

2 Related Work

In 2020, the advent of GPT-3 [2] demonstrated that large language models have outstanding linguistic capabilities and are no less effective in performing other tasks. 2021, OpenAI successfully shifted its

• *Junlin Zhu, Juanjuan Long, Yingjing Duan and Wenxuan Zhao are from Jiangnan University. E-mail: zhujunlin00@gmail.com, 8062292@qq.com, 794643458@qq.com, 870199481@qq.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

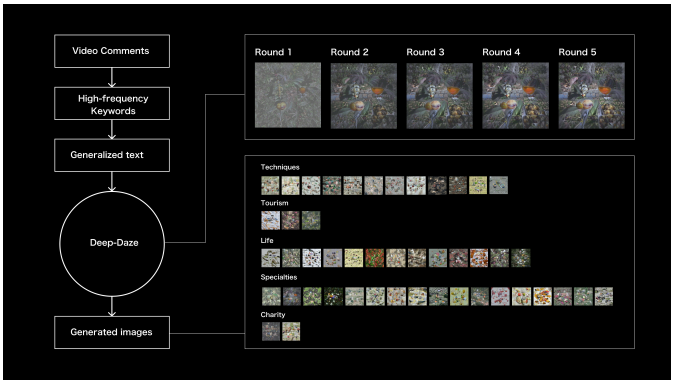

Fig. 2. Data proceesing

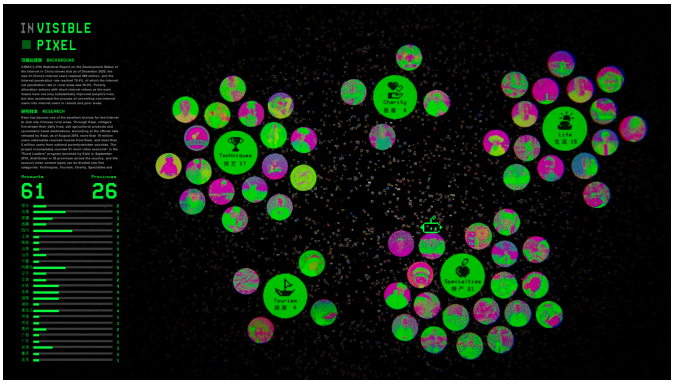

Fig. 3. ID map

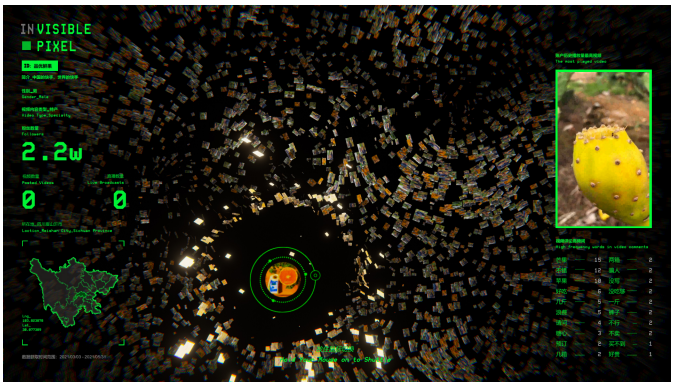

Fig. 4. Pixel tunnel

Fig. 5. Machine view

language models to the visual domain with the introduction of two models, DALL-E and CLIP [1], which implement the creation of new images from text. Among them, CLIP efficiently learns vision concepts through natural language supervision and is more efficient than the original model in accomplishing the same task. Since CLIP has trillions of parameters, there are limitations in using it, so in this project, we choose the Deep-Daze model, which is more adapted to the hardware.

Deep Daze, a text-to-image tool developed by Chinese Phil Wang in March 2021, combines a simplified version of the CLIP model with Siren [3] to convert the image content described by the text into a corresponding image. Compared with CLIP's generation of figurative images, Deep-daze generates more semantically abstract images, is visually more artistic, and is more suitable as a creative tool.

3 Data Processing

3.1 Video and image data collection

The short-video platform Kwai launched *Happy Village Leaders*, a classic case of China's poverty alleviation program. The program has discovered hundreds of potential rural entrepreneurship leaders nationwide to open video accounts on Kwai, providing them with online and offline business resources to help them improve their capabilities to increase local employment opportunities and help rural revitalization.

Since some accounts in the program are no longer updated or have less influence resulting in untraceable data, our work filters the information of 60 active accounts as data samples. We collected all short videos posted by 60 IDs up to May 28, 2021, containing 1,833 short videos, and categorized them into five content types: specialties, tourism, life, techniques, and charity. Then, we draw frames of all videos at 1-second intervals to form a dataset of images consisting of about 30,000 images.

3.2 Text data collection and Image Generation

First, we crawled the comment texts of short videos posted by 60 IDs up to May 28, 2021; then, we input these comment texts into PyCharm separately and extracted high-frequency words from them by SnowNLP to get the keywords; finally, we manually arranged these keywords into logical phrases to get the generalized text for each ID.

3.3 Image Generation

Without providing a baseline image, the generalized text of each ID is fed into Deep-Daze to go through five rounds of training (fig.2) and finally generate gifs, with each set containing about 60 frames. This process demonstrates how the machine develops its thinking and imagination: from a vague overall carving to a clear local thing. The presentation of the images is flat and intuitive, with the background reflecting the narrative environment and the narrative objects distributed in the center of the frame.

4 Interface Design and System Development

The web page has a linear structure divided into three main views: ID map (fig.3), Pixel tunnel (fig.4), and Machine view (fig.5). The system's interface is designed with elements such as color, fonts, and filters, including voiceover with machine characteristics, to create a sense of conflict of viewing the rural story with a future perspective.

The viewer first selects the account of interest in the ID map and enters a pixel tunnel consisting of video keyframes of that ID. Viewers scan the interface with their mouse to discover the avatar signal hidden behind the pixels, which maps the actual geographic location relative to the China map. After long-pressing the avatar, the ID's details are displayed, including id name, content type, location, number of followers, number of videos uploaded in the last three months, representative videos, and high-frequency words in comments. This view represents the typical perspective of our daily viewing of short videos. Long press on the avatar until it loads into the machine viewthis view presents the generalized text corresponding to the ID and the generated image from that text. Viewers can return to the main interface at any time by clicking on the icon in the upper left corner.

The work was initially developed as a web application using Unity 3D in a 4k environment, using the particle system module in Unity created particle pixel effects, with the audience interacting with the screen mainly through the mouse. It is currently migrated to the web page running on webGL, and there are problems with screen adaptation.

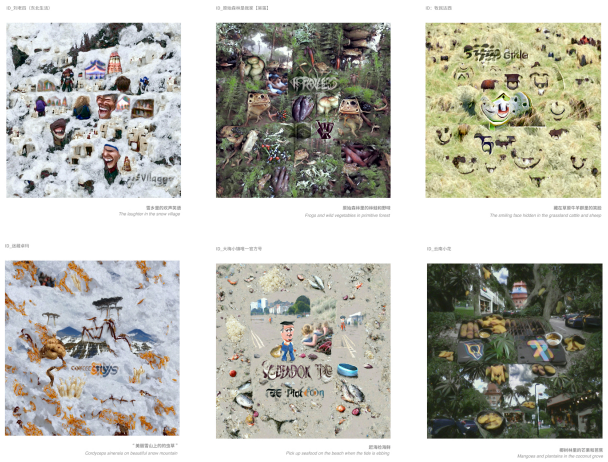

Fig. 6. Generated image samples

Therefore, it is recommended that the best viewing resolution is 1080p and full screen.

5 Conclusion

The images generated by the machine under this condition match the content type of IDs. For example, most of the generated images in the life and travel categories appear as natural scenes(fig.6), such as snow, grasslands, primitive forests, and so on, are related to their location's geographical characteristics. This work helps viewers build a visual portrait of IDs without having to browse all their videos and compare the similarities and differences between IDs by the things, scenes, words, and even colors within the images. It will also help the platform to analyze, compare and categorize the massive amount of videos posted.

Overall, this project is an artistic practice based on the latest research results in machine vision. At this stage, the machine has developed its imagination based on the dataset and make a relatively accurate visual description. The resulting images may not meet conventional aesthetic standards but offer artificial intelligence present in our future daily lives. At the same time, the high degree of objectification and automation leads to the absence of individual narrative features. Perhaps the ultimate purpose of human-machine interaction in the technological age should not be to perform calculations but to create conversations. How machines and humans create narratives together deserves further reflection and exploration.

References

[1] K. Burns, C. D. Manning, and L. Fei-Fei. Neural abstructions: Abstractions that support construction for grounded language learning. *arXiv preprint arXiv:2107.09285*, 2021.

[2] L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[3] X. Jiang, N. Y. Chen, J. I. Hong, K. Wang, L. Takayama, and J. A. Landay. Siren: Context-Aware Computing for Firefighting. 1 2004. doi: 10.1184/R1/6470429.v1