ViDA-MAN: Visual Dialog with Digital Humans

Tong Shen¹, Jiawei Zuo¹, Fan Shi¹, Jin Zhang², Liqin Jiang², Meng Chen¹, Zhengchen Zhang¹, Wei Zhang¹, Xiaodong He¹, Tao Mei¹
¹ JD AI Research, Beijing, China, ² Migu Culture Technology, Beijing, China

ABSTRACT

We demonstrate ViDA-MAN, a digital-human agent for multi-modal interaction, which offers real-time audio-visual responses to instant speech inquiries. Compared to traditional text or voice-based systems, ViDA-MAN offers human-like interactions (e.g., vivid voice, natural facial expression and body gestures). Given a speech request, the demonstration is able to provide responses with high quality videos in sub-second latency. To deliver immersive user experience, ViDA-MAN seamlessly integrates multi-modal techniques including Acoustic Speech Recognition (ASR), multi-turn dialog, Text To Speech (TTS), talking heads video generation. Backed with large knowledge base, ViDA-MAN is able to chat with users on a number of topics including chit-chat, weather, device control, News recommendations, booking hotels, as well as answering questions via structured knowledge.

CCS CONCEPTS

- Computing methodologies → Computer vision; Natural language processing; Human-centered computing → Human computer interaction (HCI)

KEYWORDS
Multimodal Interaction, Digital Human, Dialog System, Speech Recognition, Text to Speech, Talking-head Generation

ACM Reference Format:

1 INTRODUCTION

Digital humans are virtual avatars backed by Artificial Intelligence (AI), which are designed to behave like a real human. Agents powered by such systems can be applied in a wide range of scenarios such as personal assistant, customer service and News broadcasting. In this paper, we present ViDA-MAN, a multi-modal interaction system for digital humans. The system is complex by nature, integrating multimodal techniques such as ASR, TTS, dialog system, visual synthesis.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author(s).

ACM Reference Format:

2 SYSTEM ARCHITECTURE

The whole system is designed to pursue low latency and high visual fidelity, seeking intelligent and real-time interactions with a lifelike digital character. As shown in Figure 2, the system mainly consists of six modules. The system accepts human voice by an ASR module and feeds it to a dialog system. The response is further translated to realistic voice using TTS. A driving system and a rendering system are responsible for updating the appearance. A streaming service is adopted to integrate everything and encode it into media stream back to the user.
2.1 Speech Module: ASR and TTS
The speech module (ASR and TTS) acts as interpreters of the whole system. ASR has been widely studied for decades. We choose a widely-used HMM-DNN framework and use TDNN+LSTM [12] to implement the system. The TTS module is essential for human-like voice output. We adopt DIAN [13] and a speaker-specific neural vocoder LPCNet [14]. By adopting the multi-speaker duration model and acoustic model with the speaker-dependent data, the TTS system is able to clone and generate the voice of every new speaker or speaking style, which greatly enriches diversity of choices under different scenarios and can better cooperate with other modules to present digital humans.

2.2 Dialog Module: Conversation Engine
Our conversation engine is equipped with multiple dialog skills, which covers various domains including news, weather, movie, etc. As shown in Figure 2, after the ASR, the transcriptions of the user’s speech are firstly passed to a Natural Language Understanding (NLU) unit to identify the user’s intention and related entity information. The information is then formatted as the input to DST (Dialog State Tracker), which maintains the current state of the dialog. In order to coordinate different skills naturally, a hierarchical dialog policy controller [6] is adopted to select next dialogue actions. The main skills are controlled by top-level policy, whereas primitive actions are controlled by low-level policy. Finally, a model-based Natural Language Generator (NLG) [16] is involved to convert agent actions to natural language responses.

2.3 Visual Module: Driving and Rendering
The appearance of our digital human actually originates from our real human model. We first record videos of the human model speaking or making different poses. The footage is then used to reconstruct the person with intermediate representations. The role of the driving system is to produce desired representations given speech signals. Our driving system is mainly focused on the face region, i.e. making facial expressions, moving lips etc., but it can be extended to the whole body. The face is reconstructed and represented by a 3D Morphable Model (3DMM) [4, 7, 11].

The rendering system is mainly composed of a neural renderer, which converts the intermediate representation (3DMM mesh) into realistic images. We adopt a modified version of Vid2Vid [15], which is a video-to-video synthesis framework considering both spatial and temporal consistency. The model is trained by sequences of paired images, real face images and the reconstructed ones.

2.4 Realtime Streaming
Our whole system is streaming-based, therefore the streaming service unit plays an essential role in controlling the pace, encoding the frames and broadcasting the media. There are three modules involved. A play control unit is used to control the general status of the digital human. Since our driving system is mainly related to facial movement, the body frames (actions, transitions, etc.) are pre-loaded and controlled by the unit. The render system only starts to render when there are facial movements, which saves computational cost and increases scalability. The face images are then processed and fused with the body frames through a frame fusion unit and further encoded by RTC (Real-Time Communication) encoder. The frames with the synced audio content are compressed into H.264 format and broadcasted by our own RTC platform.

3 SYSTEM AND USER INTERFACE
Our dialog system is trained on multiple types of data including a knowledge graph with 180k triples, 22k entities, 20k QA pairs and million-scale NLU data, which covers topics such as celebrities, music, movies. ViDA-MAN supports tasks like enquiring weather, reserving hotels, searching news, etc. Moreover, it is also extendable to more domains in the future.

Voice and appearance are two crucial components for providing human-like interactions. Our TTS benefits a lot from the novel framework and can generate high-quality voice. On a scale from 0 to 5, our TTS system reaches 4.2 MOS (mean opinion score) in our user study, which shows superior quality. Our rendering system also benefits from our joint 2D and 3D based generation networks that produce realistic talking faces in realtime. Figure 1 shows examples of our user interface.

ViDA-MAN supports streaming of 1080p video at 25 FPS (frames per second). The overall latency is around 400 milliseconds when properly deployed. One single GPU (Tesla P40) is able to support four instances simultaneously.

ACKNOWLEDGMENTS
This work was partially supported by the National Key R&D Program of China under Grant No. 2020AAA0108600, Migu-JD Joint AI-Lab on Multimodal Interaction.
REFERENCES