Talking Avatar:
An Intelligent Mobile Application Based on Third-Party Cloud Services

Feng Ye, Hohai University, Nanjing, China & Nanjing Longyuan Micro-Electronic Company, Nanjing, China
Qian Huang, Hohai University, Nanjing, China & Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
Shengyan Wu, Hohai University, Nanjing, China & Nanjing Longyuan Micro-Electronic Company, Nanjing, China
Yong Chen, Nanjing Longyuan Micro-Electronic Company, Nanjing, China

ABSTRACT

With the booming of the mobile computing and web technology, virtual and intelligent mobile applications become increasingly popular, e.g. web computing and web-based information retrieval. However, under contemporary network conditions and web application environment, it remains a challenging problem to achieve a trade-off between algorithm complexity and hardware performance. In this article, a Talking Avatar architecture is presented based on third-party cloud services. First, the authors propose a cloud service based multi-level layered software framework, which consists of user interface layer, business logic layer and data layer. Second, human face synthesis, speech conversion and social sharing schemes are introduced to integrate third-party cloud services. Third, experimental results on Android platforms indicate that the proposed Talking Avatar can be served efficiently in terms of memory consumption as well as average response time. In addition, stronger functions are provided compared with existing methods.

KEYWORDS


1. INTRODUCTION

An increasing number of human-computer interaction requirements emerge with the advance of web and mobile computing technologies. As a result, well-known products such as Talking Tom (Outfit7 Limited, 2017) become more and more popular, especially among youngsters. And the corresponding research and development works become heated in both the academia and industry. In spite of the remarkable enthusiasms and efforts, there are still many works worthy of study and improvement. For example, according to the characteristics of the user groups, it can provide custom functionality of pronunciation and intonation; or it can provide a web-based change from real photo images into 3D cartoons, and increase the action in accordance with the user personality or emotions.

Existing studies (Lin et al., 2013; Nunes et al., 2011; Bitouk, & Nayar, 2008; Lee et al., 2010; Danihelka et al., 2011; Migliardi et al., 2012; Ezzat et al., 2004; Xie, & Liu, 2007; Wang et al., 2011; Cosatto et al., 2013; Xie et al., 2015) have shown that it is not easy to design and implement a Talking Avatar product which has rich functionality and good user experience (UE). In a nutshell,
the difficulties mainly have three aspects. Firstly, the Talking Avatar software should have good UE, and how to create vivid Avatars with a variety of decorations and how to implement the actions of the Avatars according to the characteristics, preference of the user is very important. Secondly, the hardware conditions of intelligent terminals and the influence of network environment should be considered, because resource-constrained mobile terminal and web environment is often difficult to support real-time, complex runtime requirement. Thirdly, because the Android platform is an open, fragmented ecological environment, the compatibility and integration between Talking Avatar and android platform is very complex.

As the evolution of web technologies and service computing, cloud computing and cloud services (Tsaitaris, 2014) open a new door for solving the above issues. Specific functions such as semantics understanding, face recognition and video sharing are offered by cloud service platforms (IFLYTEK Limited, 2017; Urakawa et al., 2016; Zhangtao Network Technology Limited, 2017). Based on these third-party cloud services, developers can focus on functions and quickly release satisfactory products for end users. However, many practical problems occur when the authors try to utilize cloud services, e.g. resource allocation and network bandwidth. Specific to Talking Avatar applications, the existing works cannot offer a complete Talking Avatars solution with various Avatar images and decorations, facial expressions, gestures, as well as the function of social sharing. This paper presents a Talking Avatar software architecture in which three cloud services are integrated under the Software-as-a-Service mode (Chang, 2011).

The contribution of the paper has three aspects. Firstly, the proposed solution integrating cloud services from different service providers implements a Talking Avatar product with more functions and better UE. Secondly, the authors present many local algorithms to make full use of the cloud services, e.g. similarity comparison and sub-string matching; however, it might be better to focus on the architecture due to the page limitation. Thirdly, the proposed architecture successfully moves some dirty works to the cloud service side, hence this method achieves richer functionality. The rest of this paper is organized as follows. Section 2 discusses related research works. Section 3 presents the proposed Talking Avatar architecture. Section 4 shows and analyzes experimental results. Finally, Section 5 draws the conclusion.

2. RELATED WORK

Some related works have been published in the literature, among which Talking Tom is one of the most influential products launched by the Outfit 7. Talking Tom can imitate user’s voice with a changed tone and show some predefined body movements, in addition to the social sharing function. Unfortunately, Talking Tom offers a single image and a single scenario, without any support of intelligent animation synthesis.

Lin et al. (Lin et al., 2013) proposed an interactive, multimodal and real-time 3D Talking Avatar application using cross-platform JNI and OpenGL ES. In their C/S framework, multimodal text/voice input was implemented, and facial expressions and voice animations were synthesized efficiently. However, it was lack of customizability support for Avatar images, scenarios, as well as social sharing mechanism. In addition, the proposed application was not evaluated against peer works.

Nunes et al. (Nunes et al., 2011) introduced an Avatar construction method based on basic facial photos. The authors employed the C/S framework and built an interactive and talking Avatar model. They utilized a third-party voice transferring system and focused on the recognition of mouth shapes. However, only facial features were considered in their Talking Avatar model. Furthermore, the experimental results were not evaluated against similar systems.

Bitouk and Nayar (Bitouk, & Nayar, 2008) presented a facial expression synthesis algorithm is proposed based on HMM modeling. A 3D Avatar that can speak is created with a single human photo. However, this work is not for mobile usages and supports only facial synthesis. More Avatar decorations, body gestures and facial angles can be studied for further improvement.