

摘要

真实感强的虚拟人多模式行为（语音，唇动，表情，手势等）合成除要求本身有逼真的表现外，相互之间还应该保持很好的同步和协同关系。本文主要针对这两个问题，改进并利用数据挖掘的方法开展研究，取得如下成果：

1) 在数据准备中：采用了面向 Mpeg4 标记点的语音人脸运动数据同步获取方法，可在不使用昂贵的运动跟踪设备的前提下获取按标准定义的运动数据。在数据同步分割方面，提出了一种定量的语音人脸数据同步分割方法，可较容易地获取同步数据。在数据预处理方面，采用了面向 Mpeg4 标记点的人脸动画参数生成方法，实现了从视频图象中直接提取 Mpeg4 定义的人脸动画参数。

2) 在数据特征表示与提取中：提出了面向 Mpeg-4 的人脸特征表达方法 FAPP—人脸动画参数模式，同时重点研究了通过无导师聚类以及主成分分析等方法对人脸动画参数模式的提取。在大量视频人脸运动数据的基础上，发现了 29 种基本人脸动画参数模式以及 15 个组成人脸动画参数模式的正交基。实验表明，本文提出的人脸动画参数模式及其提取方法可有效实现对人脸运动特征的提取，从而为语音人脸动画数据之间的映射和转换以及实现逼真动画打下基础。

3) 在语音人脸运动同步关联学习中：针对在语音驱动人脸动画中，如何能在考虑上下文的基础上还可实现实时的问题。本文提出了两种学习方法：一种是基于人脸动画参数模式（FAPP）的语音人脸运动神经网络映射方法；另一种是基于参数化动态转移网络（PDTN）的语音人脸运动映射方法。前者主要考虑实时性和语音的上下文关联，利用对人脸运动数据的聚类以及采用语音的前后相关帧实现了有上下文的语音到人脸动画参数模式的映射。后者在前者的基础上更进一步，不仅考虑了实时和语音的上下文，还考虑了人脸动画参数模式的上下文信息。实验表明本文提出的方法是有效的，可实现逼真的语音驱动人脸动画。

4) 在多模式行为协同韵律学习中：针对单一模式行为韵律模型学习以及多模式行为协同韵律模型获取两个问题开展研究。提出了行为合成韵律模型统一的形式化表示方法，并给出了语音韵律模型，唇动以及手势韵律模型的具体表示，同时针对汉语语音韵律模型，提出了基于多策略数据挖掘的韵律学习方法，获取了用于语音合成中韵律变化规律，取得了较好的结果。本文还给出手势与唇动，语音之间的协同韵律控制模型形式化表达。在语音，手势等韵律信息的基础上，提出了基于手语韵律与语音韵律结合的协同韵律控制模型，并应用于虚拟人多模式行为协同控制中，取得较好的结果。

5) 本文实现了两套系统及其应用示范，一种是语音驱动人脸动画系统，当给定新的语音，可以利用此模型合成出与语音同步的动画序列。第二种是文本驱动虚拟人多模式行为合成系统，当给定一个文本，可以输出具有协调一致的虚拟人多模式行为动画序列。在这两个系统的基础上，本文进行应用系统的搭建，完成面向聋人-健听人交流的对话系统以及低带宽网上虚拟人信息发布系统。前者主要通过双机翻译实现健听人和聋人之间的无障碍交流，后者主要完成基于虚拟人的 Internet 网上信息发布。两个应用系统都能较好的实现功能，满足需求。

关键词： 数据挖掘，机器学习，虚拟人合成，多模式行为，同步学习，协同韵律，语音韵律，人脸动画

Multi-modal Behaviors Data Mining for Virtual Human Synthesis

Abstract

CHEN Yiqiang (Computer Application Technology)

Directed by: GAO Wen (Professor)

To synthesize realistic virtual human multi-modal behaviors (speech, lip motion, face expression and gesture), the synchronization among these multi-modal behaviors is crucial, though the behaviors their self realistic-looking are also expected. This dissertation discusses how to apply and improve data mining method to this key problem in virtual human multi-modal behaviors synthesis. The contribution of the dissertation is as follow:

1) On data preprocessing: an mpeg-4 based labeled face feature-tracking method is adopted to obtain audio-visual synchronization data. The method not only has advantage of avoiding the expensive equipments but also has ability of obtaining accuracy data that is in accordance with mpeg-4 standard. In audio-visual synchronization data segment, a new quantitative segment method is proposed that can segment the audio-visual data more simple. In audio-visual data preprocessing, a mpeg4 labeled face feature points based face animation parameters generating method is adopted, this method explores possibility of extracting mpeg4 based face animation parameters (FAP) direct from video.

2) On data feature extraction: a new mpeg4 based visual speech data feature expression method FAPP (face animation parameter pattern) is proposed. This dissertation demonstrates on how to apply unsupervised clustering and statistic methods to FAPP extraction. Base on a large amount of audio-visual data, 29 kinds of basic FAPP that can describe face motion characteristic and 15 kinds of basic orthodoxy vector that can synthesis FAPP are obtained. The experiment shows that the proposed visual speech feature expression method can effectively realize audio-visual data mapping and vivid face animation.

3) On lip synchronization learning: Aiming towards lip synchronization problem in a speech driven face animation system, this dissertation addresses this complex many-to-many learning problem of how to design a learning model that can capture the audio and visual context information as well as real time animation. Two learning methods are proposed in this dissertation. One is FAPP based audio-to-visual neural network mapping method, the other is Parameter Dynamic Transition Network (PDTN) based audio-to-visual real time mapping method. The fore one mainly considers how to realize real time and utilize audio context information. Base on clustering method and correlation frames forward and back, the proposed method can implement the mapping from speech feature vector containing context information to face animation parameter pattern. The later one has more advantage than the fore one. It has considered not only real time and audio context information, but also utilizes the statistic context information of lip motion and expression. The experiment shows our methods are effective which can greatly improve the realistic of lip synchronization in speech driven face animation system.

4) On multi-modal behavior data synchronization learning: this dissertation addresses two

key issues on multi-modal behavior data synchronization of how to learn prosody modal for signal modal behavior synthesis and how to design the synchronization control modal for multi-modal behaviors. A uniform formalization method is proposed for behavior synthesis and the speech prosody modal, lip and gesture prosody modal are discussed in detail. A multi-strategy data-mining framework is proposed for Chinese speech prosody modal learning. The proposed framework can learn the prosody pattern for Chinese speech synthesis and the result sounds good. This dissertation also discusses about the problem of synchronization among gesture, lip motion and speech and proposes the formalization description of the synchronization control modal. Base on learned prosody parameter information of speech and gesture, a new synchronization control modal is proposed that incorporate the gesture prosody and speech prosody information and applied to control the virtual human multi-modal behavior synchronization synthesis. The experiment results sounds good given the complex of this task.

5) Two prototype systems and their applications are implemented. One is a speech driven face animation system. When new audio is given, the system can synthesize face animation sequences that synchronize with speech. The other is text to virtual human multi-modal behaviors synthesis system. When arbitrary text is given, the system can generate the virtual human multi-modal behaviors sequences with speech and lip motion and gesture synchronization. Base on these two prototype systems, we implement two kinds of application systems. One is the deaf-to-normal person dialog system, and the other is a low bandwidth networked virtual human information release system. The fore one can realize the communication between deaf and normal children through the machine translation technology. The later one can realize the information released in Internet by virtual human. Two application systems work well and satisfy the project requirement.

Keywords: Data mining, Machine learning Virtual human synthesis, Multi-modal behaviors, Synchronization, Prosody learning, Face animation, Sign language synthesis