MAC499 - Trabalho de Formatura Supervisionado

Sistema de Animação Facial Tridimensional e Síntese de Voz

Aluna: Andréa Britto Mattos

Orientador: Roberto Marcondes Cesar Jr.

IME - USP

Introdução

Este trabalho tem como objetivo complementar o projeto *Avator*, que teve início em 2006 pelo aluno Marcos Moreti[1], continuando em 2007, com a aluna Flávia Ost[2]. Nele, o usuário poderia digitar perguntas para uma face tridimensional que pronunciava as respostas. Além disso, era usada uma câmera para que a face apenas respondesse quando detectava movimento, o que foi feito utilizando a biblioteca OpenCV, de processamento de vídeo.

O Avator utilizava o programa Haptek para cuidar da animação, síntese de voz e sincronização labial do modelo. No entanto, o Haptek impunha muitas restrições para o sistema: o Avator deveria usar o Windows como plataforma, o Visual Studio como IDE e a biblioteca MFC (*Microsoft Foundation Classes*), não disponível na versão gratuita do Visual Studio.

Assim, a proposta deste projeto é a substituição do Haptek por um sistema próprio de animação e síntese de voz, utilizando um ambiente gratuito.

Interpolação de quadros-chaves

O sistema de animação utilizou a técnica de *Morph Target*, na qual um conjunto fixo de quadros-chaves são interpolados pelo computador.

Figura 1: Dadas duas poses-chaves - face de olhos abertos e de olhos fechados - o processo de interpolação gera estados intermediários, evitando transições bruscas

As poses exibidas no programa são divididas em dois grupos: *visemas* e *expressões faciais*. Um visema é o correspondente visual de um fonema, podendo corresponder a mais de um fonema. Para as expressões faciais, foi utilizada a proposta de Paul Ekman[4], na qual seis emoções puras (alegria, tristeza, raiva, medo, nojo e surpresa) podem gerar todas as outras.

Ferramentas

Para a modelar as faces e exportá-las para o sistema, foi utilizado o programa FaceGen, capaz de gerar expressões faciais, visemas e alguns movimentos não-verbais. Para a síntese de voz, foi utilizado o programa eSpeak, capaz de extrair os fonemas de um texto em português, e pronunciá-lo.

Para a animação, foi usada a engine gráfica Ogre 3D, que permite interpolar quadros-chaves e combinar *poses* com diferentes *influências*. Uma pose constitui um conjunto de valores que definem a deformação de cada vértice da malha tridimensional. A influência é a quantidade de deformação da pose. Para obter uma animação suave, foi preciso incrementar, em cada frame, a influência das poses exibidas.

Implementação

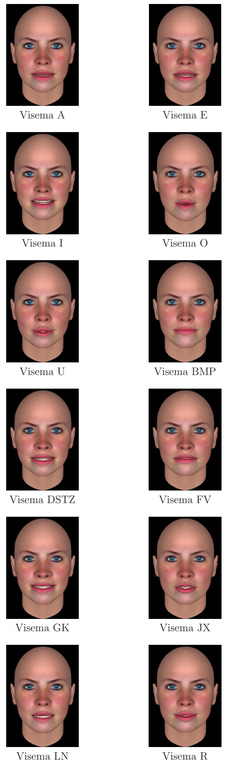
A importação dos modelos do FaceGen para o Ogre teve que respeitar as restrições de formato e animação da engine. A conversão desses modelos foi feita em dois passos: primeiro, os modelos foram convertidos para uma especificação de XML. Em seguida, foi implementado um script para modificar diretamente estes arquivos XML.

Para a sincronização labial, foi implementada uma callback que é chamada automaticamente pelo eSpeak, sempre que um buffer de áudio é produzido. Esta callback pode retornar eventos de fonema na medida em que ocorrem na fala. Estes fonemas são mapeados para o visema correspondente, que é, por fim, exibido. O mapeamento adotado é mostrado a seguir.

Mapeamento de Fonemas para Visemas			
Fonema no IPA[1]	Fonema em ASCII	Exemplo do som	Visema Correspondente
/a/	a	s**a**co	Visema A
/e/ /ɛ/	e E	s**e**co b**e**la	Visema E
/i/	i	s**i**co	Visema I
/o/ /ɔ/	o O	s**o**co s**o**zinho	Visema O
/u/	u	s**u**co	Visema U
/b/ /m/ /p/	b m p	**b**ata **m**ata **p**ata	Visema BMP
/d/ /s/ /t/ /z/	d s t z	**d**ata **s**aca **t**ata **z**aca	Visema DSTZ
/f/ /v/	f v	**f**aca **v**aca	Visema FV
/g/ /k/	g k	**g**ata **c**ata	Visema GK
/ʒ/ /ʃ/	Z S	**j**aca **ch**aga	Visema JX
/l/ /ʎ/ /n/ /ŋ/	l L n N	ga**l**o ga**lh**o **n**ata ga**nh**o	Visema LN
/ɣ/ /ɾ/	G R	ca**rr**o ca**r**o	Visema R

Tabela 1: Mapeamento de fonemas do português - no IPA e em ASCII - para grupos de visemas, exibidos nas imagens ao lado.

[1] International Phonetic Alphabet



Visão Computacional

Nesta versão, o algoritmo anterior de detecção de movimento foi substituído pela implementação do OpenCV de reconhecimento facial, para que o avatar exigisse a presença de um interlocutor humano. Foi adicionada também outra funcionalidade: mover os olhos da face conforme a posição do usuário.

Resultados

Não há Inteligência Artificial no módulo de animação, de forma que o avatar apenas repete as frases digitadas. Abaixo estão algumas *screenshots* da tela do programa. Vídeos foram também adicionados na página do projeto[3].

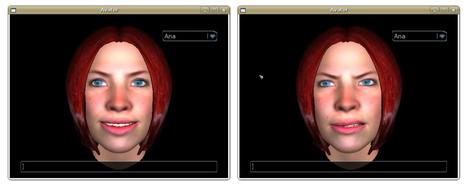


Figura 2: Avatar exibindo expressões faciais de alegria (à esquerda) e nojo (à direita)

Figura 3: Diferentes modelos que podem ser carregados no programa

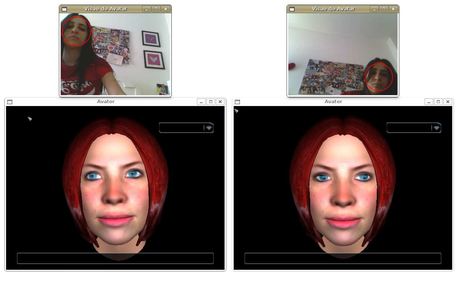


Figura 4: Movimento dos olhos do avatar seguindo a face encontrada

Referências

[1] http://www.linux.ime.usp.br/~cef/mac499-06/monografias/mpmoreti/

[2] http://www.vision.ime.usp.br/~fost/

[3] http://www.linux.ime.usp.br/~dedea/mac499/

[4] EKMAN, P.. All Emotions are Basic. Oxford University Press, Nova York, 1994.