Sistema Livre para Detecção Facial, Animação e Fala

Thales Sehn Körting[1], **Felipe Castro Silva**[1],
Rodrigo Mendes Costa[1], **Alessandro de Lima Bicho**[1]

[1]Engenharia de Computação – Fundação Universidade Federal do Rio Grande
Av. Itália, Km. 8 s/nº – 96201-900 – Rio Grande, RS, Brasil

{thales,felipecastro,rodrigoc}@vetorial.net, dmtbicho@furg.br

Resumo. *O cenário atual caracteriza-se pelo alto dinamismo, com a comunicação presencial sendo substituída pela virtual, pela praticidade e redução de custos. Surge a necessidade de criar um sistema computacional capaz de extrair movimentos da face humana para posterior simulação em uma face virtual, animada através de técnicas de Computação Gráfica. Através de vídeos adquiridos por duas* webcams*, os movimentos são extraídos utilizando técnicas de Visão Computacional e Processamento de Imagens. A face virtual é representada por uma malha de polígonos, definida por pontos (vértices) no espaço 3D, conectados por arestas. Utilizando a linguagem de programação JAVA, pretende-se propor à comunidade do Software Livre um sistema de detecção e animação facial para a fala brasileira, em conjunto com um sistema Texto-Fala já desenvolvido.*

Abstract. *Actual scenario has high dynamism, with presential communication being changed by virtual, for practicity and low costs reasons. So, it grows the aim of creating a computational system to extract movement of a human face to simulate in a virtual face, animated through Computer Graphics techiniques. Using two videos from webcams, the movements are extracted by Computational Vision and Image Processing techniques. The virtual face is represented by a polygon mesh, and a set of vertices in the 3D space, connected through edges. Using JAVA programming language, the aim of this project is show to the Free Software Community a system for Facial Detection, Animation and Speech, adapted to the Brazilian case.*

1. Introdução

O advento da Computação Gráfica e a rápida emersão de *hardwares* próprios para a execução de aplicações tridimensionais, permitiram novos estudos sobre a modelagem de estruturas complexas. Neste contexto, surgiu a proposta de modelar a face humana, de forma a reconhecer e simular seus movimentos em um ambiente virtual, através do desenvolvimento de um *framework* que possa auxiliar inúmeras ferramentas, tornando a interação com o computador mais amigável.

Para que este objetivo seja alcançado, inicialmente é necessária a extração dos movimentos de um modelo de face humana. Utilizando técnicas de Visão Computacional e Processamento de Imagens, pode-se coletar as características faciais[1] estáticas e dinâmicas necessárias de um modelo humano real. Estas características são representadas através de pontos do espaço tridimensional, possibilitando a animação de uma face virtual do modelo humano analisado.

O propósito deste artigo é o de divulgar um sistema de animação facial dotado de fala, particularmente desenvolvido para o caso brasileiro (português). Esse sistema é composto de diversos módulos: um módulo de extração dos movimentos de um personagem real, um

[1]Características faciais são os elementos que compõem a face, como sobrancelhas, olhos, nariz, lábios, queixo, entre outros.

módulo de animação da face com base no treinamento pré-realizado, e um módulo de fala, através de um Sistema conversor Texto-Fala (TF).

As próximas seções apresentarão cada um dos módulos já citados, os quais já estão implementados e disponíveis como ferramentas livres. A seqüência do artigo inicia-se pela extração dos movimentos da face (Seção 2), seguida da animação facial (Seção 3) e posteriormente da fala brasileira (Seção 4). Após, considerações sobre os módulos implementados são apresentados na Seção 5 e, por último, as conclusões e propostas de integração são apresentadas na Seção 6.

2. Extração dos movimentos faciais

Para extrair os pontos característicos da face foi desenvolvida uma metodologia em camadas, onde cada etapa anterior subsidia as posteriores. A seguir, descreve-se cada uma das etapas que compõem o processo de extração dos movimentos faciais, como mostra a Figura 1.

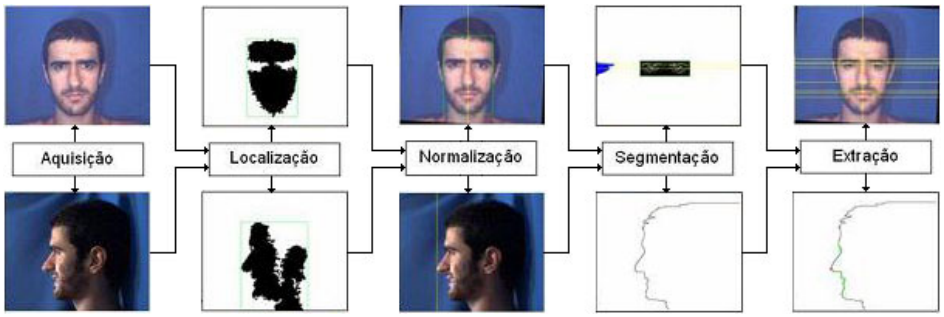


Figura 1. Etapas do processo de extração de características para cada *frame*.

2.1. Aquisição

Para adquirir os vídeos de um modelo humano, utilizou-se câmeras CCD[2] USB posisionadas frontal e lateralmente ao modelo humano. Conhecidas como *webcams*, essas câmeras apresentam uma resolução muito baixa (352 x 288 *pixels*) em comparação com as câmeras tradicionais, além de serem muito sensíveis as variações de luminosidade. O *framerate* médio conseguido aproximou-se de 10 *fps*.

2.2. Localização

Após a aquisição dos vídeos, deve-se localizar a face em cada imagem (ou *frame*), antes de extrair cada característica facial em particular. Isto facilitará o trabalho, pois a procura por elementos não será em toda imagem, mas sim em um espaço reduzido. A utilização da cor da pele para determinar a localização da face é uma subárea dos métodos baseados em características invariantes[3]. Para localizar os possíveis pontos da pele foi desenvoldida uma técnica que une as abordagens de [Buhiyan et al. 2003] e [Sheng 2003].

Os métodos baseados na cor da pele para localização de uma face geralmente apresentam ruído devido ou a um fundo complexo ou a objetos (que possuem cor semelhante a da pele humana) que estejam na imagem. Para resolver este problema, utilizamos um algoritmo de detecção de objetos descrito em [Silva et al. 2006].

[2]CCD (*Charge-Coupled Device*) é um sensor para a gravação de imagens formado por um circuito integrado contendo um conjunto de capacitores ligados (acoplados) [Wikipedia 2005].

[3]Baseia-se na premissa de que a cor pura da pele não varia durante o vídeo.

2.3. Normalização

O processo de normalização é responsável pela restauração e realce de imagens. O problema das rotações naturais da face foi tratado utilizando o ângulo extraído pelo método hídrido de estimação do ângulo [Silva et al. 2006]. Posteriormente, foi aplicada uma filtragem passa-baixa, necessária para a redução de ruídos inerente à captura de quadros de vídeo CCD de baixa resolução.

2.4. Segmentação

Uma vez normalizada a imagem e localizada a região que compreende a face, a próxima tarefa a ser realizada é a de segmentar suas características faciais internas. Neste ponto, a informação sobre o eixo de simetria da face já é conhecida, podendo ser útil para delimitar as regiões que contém cada elemento da face. Os algoritmos de Sobel [Gonzalez and Woods 2000] e de Canny [Canny 1986] foram empregados nesta fase.

2.5. Extração

Para a extração de pontos nas imagens frontais da face utilizou-se o método das projeções integrais. Segundo [Seara 1998], uma borda é definida por uma mudança no nível de cinza, quando ocorre uma descontinuidade na intensidade, ou quando o gradiente da imagem tem uma variação abrupta. Um operador de derivada é sensível a estas mudanças, operando como um detector de bordas. Onde os valores da intensidade da imagem possuírem derivada como um ponto de máximo, ter-se-á marcado suas bordas. Para as extração dos pontos das imagens laterais, utilizou uma análise sobre o perfil do modelo humano. Tendo-se a linha de perfil, uma análise de picos e vales consegue detectar pontos característicos.

3. Animação Facial

Dada uma malha de polígonos, deve-se realizar a associação entre os pontos de controle do padrão MPEG-4 [Pandzic and Forchheimer 2003] e o modelo 3D. Tendo estabelecidos os pontos de controle na malha facial, aplica-se a ela uma dinâmica de movimento. Essa dinâmica contém os movimentos necessários para a fala, os quais estarão associados aos arquivos gerados pelo sistema Texto-Fala.

3.1. Técnicas de Animação Facial

Duas técnicas foram empregadas na implementação deste módulo: a de Waters, que simula músculos virtuais na malha facial; e a das funções de base radial (*Radial Basis Functions* – RBF's), sendo descritas a seguir. O modelo de Waters é empregado para a simulação de emoções na face (*e.g.* alegria, dúvida, raiva, etc.) e as RBF's são utilizadas para a animação dos pontos utilizados na produção da fala, principalmente aqueles situados nos lábios.

3.1.1. O modelo de Waters

Também conhecido por *Muscle Based Animation*, foi primeiramente implementado por Waters, em 1987 [Waters 1987]. Esse modelo se baseia na construção de músculos virtuais, interagindo com a malha representativa da face. Nessa estratégia, cada músculo é representado como um vetor (presente em algum ponto da face) que, ao ser aplicado, desloca na sua direção os demais vértices das regiões vizinhas, de acordo com sua zona de influência, definida por uma série de parâmetros. Maiores detalhes sobre esta técnica estão disponíveis em [Korting et al. 2005b].

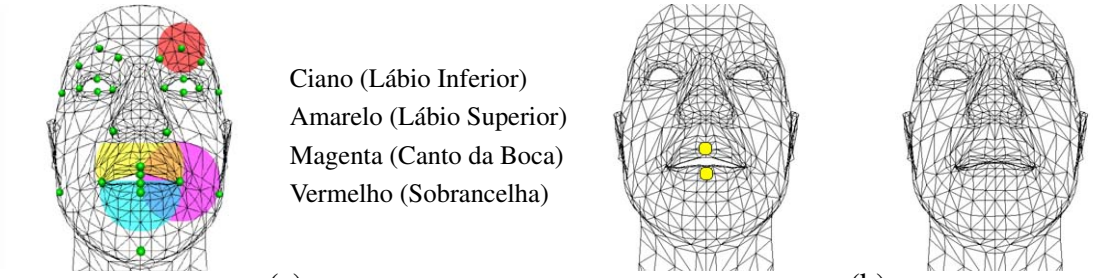


Figura 2. (a) Regiões de atuação de alguns pontos de controle (b) Movimento de pontos de controle da boca aplicados à vizinhança.

3.1.2. *Radial Basis Functions*

O uso das RBF's proporciona deformações suavizadas com um comportamento facilmente controlável. RBF's são funções que interpolam dados fornecidos, produzindo resultados contínuos. A geometria dos pontos de controle não necessita restrições, implicando em uma distribuição dos mesmos de forma esparsa e irregular. Além disso, tal comportamento pode ser empregado para se chegar aos requisitos desejados em uma animação [Wirth 2000].

De um modo geral, o funcionamento das RBF's aplicado à animação facial é dado da seguinte maneira: pontos de controle são definidos na superfície facial em um estado inicial, com a face relaxada. Para se obter a animação de alguma característica (*e.g.* canto esquerdo da boca), é atribuído um ponto de controle (chamado *landmark*) para essa região, e se movimenta tal ponto para uma região alvo. Após, são calculadas as posições dos pontos que estiverem na vizinhança desse *landmark*.

Para realizar a animação, temos dois estados: um inicial ou relaxado, e outro final, contendo a movimentação do(s) ponto(s) de controle e da sua região de influência. Com esses estados (ou *frames*), produz-se a animação através de uma técnica denominada *morphing*[4].

Estabelecidos os pontos de controle, resta saber suas respectivas regiões de influência, tendo em vista que determinados pontos apresentam regiões restritas (não são regiões circunscritas por um determinado raio). Na Figura 2.a são apresentadas as principais regiões de influência dos pontos de controle. Na Figura 2.b, a aplicação das RBF's serve para simular o movimento da boca, apenas com o deslocamento de dois pontos de controle em destaque. Mais exemplos, bem como diversas regras para animação facial podem ser encontrados em [Korting et al. 2006].

4. Texto-Fala

Esta seção descreve o terceiro módulo envolvido no sistema, responsável pela fala humana, implementado por meio do sistema Texto-Fala[5] (TF). A síntese concatenativa foi utilizada no processo de síntese de voz. Essa técnica processa uma pequena quantidade de dados (ou difones[6]) para gerar o som. Assim, a síntese do som é feita através da concatenação desses sons, manipulados de modo que possam se "encaixar".

[4]Algoritmos que realizam um alinhamento geométrico pela interpolação das cores afim de produzir uma transição suavizada e realista entre dois estados.

[5]Texto-Fala é um sistema computacional que deve ser capaz de ler qualquer texto [Dutoit 1997]

[6]Difones são as unidades da fala que iniciam no meio de um estado estável de um fonema e terminam na metade do seguinte [Dutoit et al. 1998]

O sistema TF é composto de dois módulos principais: o Processador de Linguagem Natural e o Processador de Sinais Digitais [Dutoit 1997]. O primeiro é o responsável pela produção da transcrição fonética do texto a ser lido, juntamente com a entonação e o ritmo mais adequados, através da geração de uma informação simbólica. O segundo componente é responsável pela transformação da informação gerada pelo módulo anterior em fala.

Cabe ao processador de linguagem natural o processamento e a análise dos textos de entrada, além da realização do casamento de padrões com regras, abreviaturas, letras ou números. Os módulos internos do sistema Texto-Fala estão descritos em [Korting et al. 2005a]. A conversão é realizada de maneira implícita, através do armazenamento prévio de exemplos de transições fonéticas em uma base de dados de fala, também conhecidos como unidades acústicas [Dutoit 1997]. As unidades acústicas são então concatenadas através do método utilizado nesse sistema, o da síntese concatenativa. Nesse ponto, é utilizado o sistema MBROLA (*Multi Band Resynthesis OverLap Add*) para a manipulação desta base de unidades acústicas necessária ao processador de Sinais Digitais.

Foi levado em consideração a geração de prosódia (pronúncia regular da palavras com a devida acentuação), sendo esta uma das mais difíceis tarefas no estudo de síntese de voz. A falta de um modelo explícito de fala (característica emocional adaptada ao contexto), na maioria das estratégias de síntese concatenativa, limita a utilidade de muitos sistemas atualmente desenvolvidos à tarefa restrita de uma fala meramente neutra.

Esse sistema será o responsável pela geração dos sons, através de um texto em português anteriormente fornecido. Como já divide as palavras nos fonemas mais adequados para a conversão, estes serão utilizados pelo módulo de animação. Esses movimentos, realizados para a geração dos sons, denominam-se visemas.

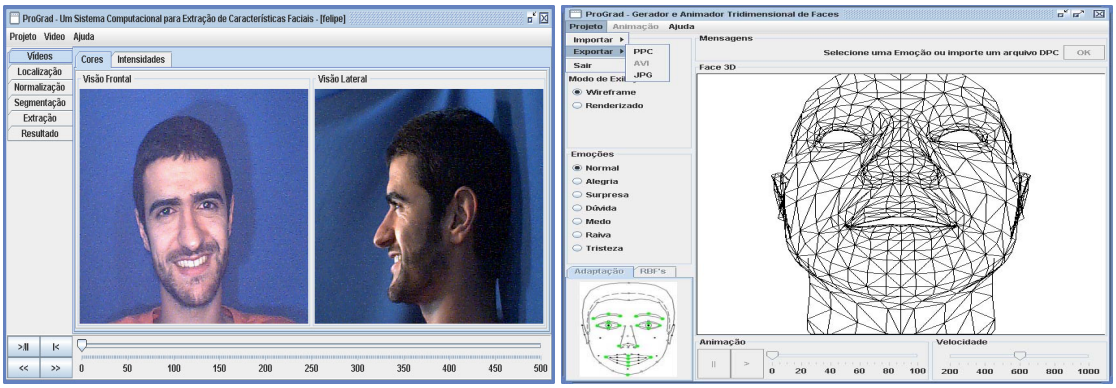
5. Considerações finais

Os módulos descritos nas seções anteriores foram implementados utilizando-se a linguagem JAVA, sendo esta decisão embasada nos recursos providos por esta linguagem. Um dos objetivos deste artigo, além de divulgar as implementações de diversas técnicas da área de processamento gráfico e de voz, é propor à comunidade de Software Livre a continuidade de desenvolvimento destes módulos, com o intuito de integrá-los em um único sistema denominado pelos autores como *Facial DAS - Detection, Animation and Speech*. Cabe salientar que todos os módulos estão funcionais, permitindo que os recursos disponibilizados sejam facilmente integrados em outros sistemas computacionais.

Na Figura 3.a há um *screenshot* da interface do módulo responsável pela extração das características faciais. Por meio da definição de um projeto na interface, o usuário importa pares de vídeos contendo imagens frontais e laterais de modelos humanos. Após a inserção dos vídeos, a extração das características faciais pode ser realizada, sendo possível a visualização de cada uma das etapas do processo em separado. Um vídeo demonstrando a interatividade da implementação está disponível em `http://facialdas.sf.net/video_detection.html`.

Na Figura 3.b há um *screenshot* da interface do módulo responsável pela animação facial. Ao se iniciar um novo processo, é necessário importar uma malha de polígonos contida em arquivos do formato OBJ, adaptando os pontos da face à malha lida. Após, realiza-se a associação entre os pontos de controle da imagem 2D (canto inferior esquerdo na interface) ao modelo 3D. Finalmente, pode-se utilizar uma "emoção" padrão ou importar uma dinâmica de movimento extraída de vídeos. Um vídeo demonstrando a interatividade da implementação está disponível em `http://facialdas.sf.net/video_animation.html`.

Em relação ao módulo texto-fala, é possível verificar a implementação realizada através do link `http://fttss.sf.net/`. Todos os módulos, assim como a documentação referente, estão disponíveis em `http://facialdas.sf.net/`.



(a) (b)

Figura 3. Interface em (a) do módulo responsável pela extração dos movimentos faciais e em (b) do módulo responsável pela animação facial.

6. Conclusão

Um sistema de detecção e animação facial foi apresentado, juntamente com um sistema Texto-Fala. Aspectos práticos de implementação foram mostrados, divulgando as técnicas utilizadas e os módulos desenvolvidos à comunidade científica. Como já mencionado, o intuito deste artigo é o de divulgar a existência destes módulos e o desenvolvimento de um *framework* capaz de auxiliar no reconhecimento dos movimentos de um usuário e simulá-lo, juntamente com a voz, em um computador.

Embora todos os três módulos do sistema já estejam em funcionamento, a adaptação dos subsistemas descritos ainda não está completa. Para isso, conta-se com a contribuição da comunidade de Software Livre, tendo em vista um sistema que muito pode contribuir para a comunicação virtual.

Referências

Buhiyan, M. A. A., Ampornaramveth, V., and Ueno, S. Y. H. (2003). Face Detection and Facial Feature Localization for Human-machine Interface. *NII Journal*, 5.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698.

Dutoit, T. (1997). High-quality text-to-speech synthesis: An overview. *Electrical and electronics engineering*, 17(1):25–36.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and Vrecken, O. V. D. (1998). The mbrola project: Towards a set of high quality speech synthesizers free of use for non-commercial purposes. In *Proc. ICSLP'96*, pages 1393–1396, Philadelphia, USA.

Gonzalez, R. C. and Woods, R. E. (2000). *Processamento de Imagens Digitais*. Editora Edgard Blücher Ltda.

Korting, T. S., da Silva, F. C., and Costa, R. M. (2005a). Um Sistema Texto-Fala Livre. In *VI Workshop de Software Livre/WSL'05, FISL6.0*.

Korting, T. S., da Silva, F. C., Costa, R. M., de Lima Bicho, A., and da Costa Botelho, S. S. (2005b). Um estudo sobre a animação tridimensional de faces. In *Workshop de Iniciação Científica, XVIII SIBGRAPI*.

Korting, T. S., Silva, F. C., Costa, R. M., and de Lima Bicho, A. (2006). Animação facial baseada em funções de base radial. In *Workshop de Iniciação Científica, XIX SIBGRAPI*.

Pandzic, I. and Forchheimer, R. (2003). *MPEG-4 Facial Animation*. John Wiley & Sons.

Seara, D. M. (1998). Algoritmos para Detecção de Bordas. Disponível em `http://www.inf.ufsc.br/~visao/1998/seara/`.

Sheng, Y. (2003). Fast and Automatic Facial Feature Extraction for 3-D Model-Based Video Coding. *Research Excellence Awards Competition 2003*.

Silva, F. C., Costa, R. M., Korting, T. S., and de Lima Bicho, A. (2006). Estimando a angulação da face humana através de um método híbrido (elipse-simetria). In *Workshop de Iniciação Científica, XIX SIBGRAPI*.

Waters, K. (1987). A muscle model for animation three-dimensional facial expression. In *SIGGRAPH '87:*, pages 17–24, New York, NY, USA. ACM Press.

Wikipedia (2005). CCD. Disponível em `http://pt.wikipedia.org/wiki/CCD`. acessado em 31/12/2005.

Wirth, M. A. (2000). *A Nonrigid approach to Medial Image Registration: Matching Images of the Breast*. PhD thesis, RMIT University, Melbourne, Victory. Australia.