Laboratório VISGRAF
Instituto de Matemática Pura e Aplicada

Modelos Avancados de Animacao Facial: O Estado da Arte

Paula Salgado Lucena Rodrigues, Bruno Feijo, Luiz Velho.

Technical Report TR-2005-02 Relatório Técnico

April - 2005 - Abril

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Modelos Avançados de Animação Facial
– O Estado da Arte –

por

Paula Salgado Lucena Rodrigues

Orientadores: Prof. Bruno Feijó e Prof. Luiz Velho

Abril/2005.

Sumário

Lista de Figuras

1. Introdução

Esta monografia tem como propósito apresentar um estudo do estado da arte da pesquisa em animação facial, descrevendo e comparando os principais trabalhos que vêm sendo desenvolvidos nessa área da computação gráfica.

O material de base para esta monografia consiste de uma coletânea de artigos que foram publicados nos anos de 2003 e 2004. No entanto, alguns artigos de anos anteriores também foram incluídos devido a sua importância ou para o entendimento e complementação de alguns dos trabalhos expostos.

Com o objetivo de apresentar um trabalho mais completo, são apresentados, ainda na "Introdução" deste documento, alguns conceitos básicos sobre animação.

Basicamente, para cada artigo, é feita uma apresentação sumária do que o artigo propõe. Em seguida, é apresentado um resumo do artigo completo, procurando sempre respeitar as seções desenvolvidas no próprio artigo. Por fim, após a apresentação das conclusões do artigo, há uma subseção intitulada "contribuições pessoais" onde é feita uma análise do artigo em questão.

Após a apresentação e análise de cada um dos artigos, uma seção de conclusões encerra esta monografia oferecendo uma comparação entre os trabalhos e apontando os pontos de pesquisa em aberto.

1.1. MODELANDO FACES

O desenvolvimento de um modelo facial envolve determinar sua descrição geométrica e sua capacidade de animação. Adicionalmente, é correto pensar que também está envolvida nessa etapa a representação de atributos adicionais para a face, tais como superfícies coloridas e a utilização de textura. O foco desta seção é apresentar, de forma resumida, as principais técnicas de modelagem facial "animáveis". Isso significa que "faces estáticas" não são interessantes, as técnicas de modelagem interessantes para este documento são as que possibilitam a animação do modelo facial definido.

O objetivo das várias técnicas de animação, que serão apresentadas na Seção 1.2, é controlar as faces modeladas. Esse controle precisa ser feito de tal forma que as superfícies faciais renderizadas tenham o formato, as cores e as texturas (as duas últimas quando aplicadas) desejadas em cada quadro da seqüência animada [Parke1996]. É também importante que a modelagem favoreça a obtenção de uma animação e uma renderização (*rendering*) eficientes.

1.1.1. Representações Volumétricas

Uma das abordagens de modelar faces é através da utilização das várias técnicas de representação volumétricas. Isso inclui *Constructive Solid Geometric* (CSG), *arrays* de elementos de volume (*voxels*) e elementos de volumes agregados, tais como *octrees* [Parke1996].

CSG é utilizada com sucesso em um número grande de sistemas mecânicos "*computer-aided"*. Nesses sistemas, os objetos de interesse são representados utilizando

construções de conjuntos booleanos com formato matemático regular relativamente simples, como planos, cilindros e esferas.

Normalmente, faces realistas não são representadas dessa forma. Com isso, CSG não é uma técnica de base geométrica popular para faces. Por outro lado, é possível imaginar um personagem tridimensional com estilo *cartoon* modelado (construído) a partir da técnica CSG.

Representação de elemento de volume (*voxel*) é a forma preferida de descrever estruturas anatômicas em imagens médicas. O ponto negativo é que a representação por *voxel* tipicamente requer uma quantidade grande de memória. Sendo assim, "modelos diretos" de *voxel* não são atualmente usados para animação facial [Parke1996]. No entanto, técnicas, como os algoritmos de *Marching Cubs*, podem ser usadas para extrair modelos geométricos de superfícies anatômicas a partir dos dados do *voxel*. A animação pode então ser feita utilizando modelos de superfícies extraídas.

1.1.2. Representações de Superfícies

Superfícies primitivas e estruturas são, atualmente, a base geométrica preferida para modelos faciais [Parke1996] [Zhang2003]. As estruturas de superfície utilizadas permitem formatos de superfície e mudanças nos formatos quando necessário para as várias conformidades e expressões faciais. Possíveis técnicas de descrição de superfícies incluem *superfícies implícitas*, *superfícies paramétricas* e *superfícies poligonais*. Superfícies paramétricas incluem *Bézier bivariável*, *Catmull-Rom*, *Beta-spline*, *B-Spline*, *B-Spline hierárquica* e superfícies *NURBS*. Superfícies poligonais incluem malhas regulares poligonais e redes de polígonos arbitrárias.

1.1.3. Novas Faces a partir de Faces Existentes

Várias abordagens, como será observado ao longo dos artigos apresentados, têm sido propostas para a criação de novas faces baseando-se em faces existentes. Esses processos incluem interpolação entre faces existentes, aplicar deformação a faces existentes e transformar uma face canônica em faces de indivíduos específicos [Parke1996].

1.2. ANIMAÇÃO FACIAL

A Seção 1.1 destinou-se a apresentar, de forma simples, as técnicas usadas no processo de modelagem facial. Uma vez que a face foi modelada, o passo seguinte consiste de sua animação. Esta seção destina-se a apresentar algumas das técnicas principais de animação facial.

Existem ao menos cinco abordagens fundamentais para animação facial. Essas abordagens são: interpolação (*interpolation*), baseada em performance (*performance-driven*), parametrização direta (*direct parametrization*), animação baseada em pseudo-músculos (*pseudomuscle-based animation*) e animação baseada em músculos (*muscle-based animation*).

1.2.1. Interpolação

A interpolação é uma das formas de manipular superfícies flexíveis, tais como superfícies utilizadas nos modelos faciais. A interpolação é, provavelmente, a técnica mais usada na animação facial, sendo também uma das técnicas mais simples. Por exemplo, no caso unidimensional são dados dois valores e deseja-se determinar um valor intermediário, onde esse valor intermediário é especificado por um coeficiente de interpolação fracionária α:

$$value = \alpha(value_1) + (1.0 - \alpha)(value_2)$$

com $0.0 < \alpha < 1.0$.

Esse conceito básico é facilmente expandido para mais do que uma dimensão, aplicando-se esse mesmo procedimento em cada uma das dimensões. A idéia pode ser generalizada para superfícies poligonais, aplicando-se esse esquema em cada vértice que define a superfície. Cada vértice possui duas posições tridimensionais associadas com ele. Formas intermediárias da superfície são alcançadas através da interpolação de cada vértice entre suas posições extremas.

1.2.1.a. Interpolação de Expressão Chave

Entre os mais antigos e ainda mais fortemente esquemas utilizados para implementação e controle de animação facial está a interpolação de poses-chave de expressões faciais (*key expression pose*) [Parke1996]. A idéia consiste em coletar dados geométricos que descrevem a face em ao menos duas poses de expressões diferentes. Depois, um parâmetro de controle, o coeficiente de interpolação, é utilizado como função do tempo para "trocar" a face de uma expressão para a outra. Uma suposição básica que pode ser feita para a interpolação das superfícies faciais é que uma única topologia facial possa ser usada para cada superfície. Se a topologia da superfície é fixa, a manipulação do formato da superfície consiste, basicamente, da manipulação das posições dos vértices.

Modificar a face de uma expressão para outra é um problema de mover cada ponto de controle da superfície (os vértices) de uma distância pequena em quadros sucessivos.

A interpolação de expressões pode ser estendida de várias formas. Algumas dessas maneiras estão citadas abaixo:

- **Interpolação Bilinear de Expressão:** se quatro expressões estão disponíveis, então dois parâmetros de interpolação podem ser utilizados para gerar uma expressão que seja a combinação bilinear das quatro poses chaves. Se oito expressões estão disponíveis, são necessários três parâmetros de interpolação para gerar uma expressão combinada trilinear.
- **Interpolação *n*-Dimensional de Expressão:** Quatro parâmetros de interpolação e dezesseis expressões chaves permitem uma combinação no espaço de interpolação quatro-dimensional. Interpolações em um espaço de expressão de maiores dimensões é possível, mas provavelmente não é útil para o animador porque esses espaços de dimensão muito grande não são muito intuitivos.

- **Interpolação *Pairwise* de Expressão:** uma outra forma de explorar múltiplas poses de expressões é permitindo uma seleção *pairwise* das poses a partir de uma biblioteca de expressões. Dessa forma, um único parâmetro de interpolação é utilizado para combinar as poses selecionadas.
- **Interpolação por Região Facial:** uma outra extensão bastante útil para interpolação de expressões é dividir a face em um número de regiões independentes. Valores separados de interpolação são aplicados a cada região. Um exemplo dessa abordagem é dividir a face entre região superior e região inferior. Normalmente, a região superior é utilizada para as expressões emocionais, enquanto a região inferior é utilizada para expressões da fala.
- **Interpolação Não-Linear:** como a face é governada por leis físicas, seus movimentos não são lineares, pelo contrário, eles tendem a acelerar e desacelerar. Na interpolação linear, o valor interpolado é computado usando uma função linear de dois pontos finais (pontos extremos). Porém, não existe nenhuma restrição para o uso de um coeficiente de interpolação α que seja determinado como função do tempo de quadro da animação. Esse valor α pode ser uma função não-linear do tempo. Por exemplo, [Parke1996] encontrou funções baseadas em cossenos de intervalos de tempo fracionários que são muito úteis para as aproximações de aceleração e desaceleração.

Por fim, o esquema de interpolação possui limitações. Primeiro, o intervalo de controle das expressões está diretamente relacionado com o número e a disparidade das poses das expressões disponíveis. Uma expressão fora dos limites do conjunto de poses chaves é inatingível, exceto talvez por extrapolação, que é uma abordagem bastante arriscada. Uma segunda limitação está no fato de que cada pose chave necessita de uma coleção explícita do dado geométrico ou do esforço da geração dos dados. Para um conjunto muito grande de poses isso se torna uma tarefa difícil.

1.2.2. Animação Baseada em Performance

A técnica de animação baseada em performance (ou dirigida por performance, como também é chamada) consiste em utilizar informações derivadas por uma medida de ações reais humanas para dirigirem personagens sintéticos. Animação baseada em performance normalmente faz uso de dispositivos de entrada, tais como, *data gloves* e *laser- or – video-based motion capture* [Parke1996]. Duas abordagens são discutidas aqui para animação baseada em performance: mapeamento de expressão (*expression mapping*) e transmissão da pessoa baseada no modelo (*model-based persona transmission*).

O mapeamento de expressão tem como passo inicial à aquisição de um número digitalizado de poses de expressões e de fonemas tirados de uma pessoa real. Esse passo pode, por exemplo, ser feito através de uma câmera fotográfica digital. O passo seguinte consiste em fazer uma correspondência entre a expressão neutra da face real e a expressão neutra do personagem a ser animado. Como o número de pontos que definem a face do personagem normalmente é maior que o número de pontos definidos na face real humana, é utilizada uma correspondência *um-para-n*. É importante salientar que cada ponto na face do personagem deve ter apenas um único ponto correspondente na face

real. Por fim, uma vez que essa correspondência foi definida, as expressões do personagem são computadas utilizando-se uma função que mapeia as expressões da face real para o personagem.

Já na abordagem transmissão da pessoa baseada no modelo, Parker [Parke1996] sugere que uma representação sintética convincente da face de uma pessoa pode ser usada para criar um videofone sintético. Isso é possível se técnicas de análise poderosas forem utilizadas/desenvolvidas para fazer o casamento (*matching*) dos movimentos de uma imagem sintética com os movimentos da pessoa real. Tal sistema pode operar numa taxa muito baixa de canal de dados, como a de uma rede de telefonia padrão.

Os aspectos da análise e da síntese da imagem são igualmente importantes nessa abordagem. A análise de imagem precisa extrair automaticamente todos os parâmetros relevantes da imagem origem em tempo real. Essa análise inclui *tracking* dos movimentos de cabeça, identificação das características faciais e identificação dos formatos característicos. Essa informação extraída é então transmitida para o sistema remoto de síntese de imagem para produzir as imagens faciais sintéticas correspondentes. É importante ressaltar que o sucesso dessa abordagem depende da geração em tempo real de faces sintéticas convincentes.

1.2.3. Modelos Diretos Parametrizados

Motivado pelas dificuldades associadas com a animação *key-pose*, Parke [Parke1996] desenvolveu uma técnica de animação de modelos diretamente parametrizados. O intuito dessa proposta é criar um modelo encapsulado que gere um intervalo grande de faces e de expressões faciais baseado no menor conjunto de parâmetros de controle possível. O objetivo é permitir que tanto as expressões faciais quanto à conformidade facial sejam controladas por um conjunto de valores dos parâmetros.

O ideal seria ter um modelo que permitisse todas as faces possíveis com quaisquer expressões para ser especificada através da seleção do conjunto de valores apropriados dos parâmetros. No entanto, os modelos criados até o momento, segundo Parke em [Parke1996] e pelos trabalhos atuais estudados como [Chai2003] e [Cao2003] , estão ainda distante do modelo tido como ideal, mas eles já permitem um grande número de expressões.

O desafio é de determinar um bom conjunto de parâmetros de controle e implementar um modelo que faça uso desses parâmetros a fim de gerar o intervalo desejado de faces e de expressões. Os parâmetros de controle, basicamente, devem atender (incluir) os seguintes elementos:

- **Expressão:** abertura das pálpebras, arco da sobrancelha, separação da sobrancelha, rotação da mandíbula, largura da boca, expressão da boca, posição do lábio superior, posição dos cantos da boca, pupilas etc.
- **Conformidade:** largura das mandíbulas, formato da testa, largura e tamanho do nariz, formato do queixo, formato do pescoço, tamanho e separação dos olhos, proporções das regiões da face e proporções gerais da face.

Estudos em [Parke1996] levaram a deduzir que cerca de 10 parâmetros de expressões permitem que o animador especifique e controle um grande intervalo de

expressões faciais. Cerca de 20 parâmetros são usados para controlar o intervalo limitado de conformidade das expressões.

1.2.4. Animação Baseada em Pseudo-Músculos

A interação complexa entre os tecidos, músculos e ossos da face e apenas entre os próprios músculos resulta no que normalmente é chamado de "expressões faciais". É evidente que essas interações produzem um enorme número de combinações de movimentos.

A idéia para a abordagem "animação baseada em pseudo-músculo" é não exatamente simular a anatomia detalhada da face, e sim, desenvolver modelos que apenas com o controle de alguns parâmetros seja possível simular as ações básicas dos músculos da face. Exemplos do uso de animação baseada em pseudo-músculos é o trabalho desenvolvido em [Thalmann1988] e em [Kalra1992].

1.2.5. Animação Facial Baseada em Músculos

A anatomia detalhada da cabeça e da face é uma complexa junção de ossos, cartilagens, músculos, nervos, vasos sangüíneos, glândulas, tecido adiposo, tecidos de conectividade, pele, cabelo etc. [Parke1996]. Até o presente instante, segundo afirmação de [Parke1996], nenhum modelo facial baseado nessa detalhada anatomia foi reportado. Por outro lado, muitos modelos foram desenvolvidos baseados em modelos simplificados usando a estrutura facial dos ossos, músculos, tecidos de conectividade e pele. Esses modelos possuem a habilidade de manipular as expressões faciais através de uma simulação das características dos músculos e dos tecidos faciais.

Platt e Badler [Platt1981] desenvolveram um modelo dinâmico de face onde os vértices poligonais da superfícies da face (a pele) eram elasticamente interconectados através de molas que foram modeladas. Esses vértices também eram conectados à estrutura de ossos interior ao modelo usando simulação de músculos. Esses "músculos" tinham propriedades elásticas e podiam gerar forças de contração. As expressões da face eram manipuladas através da aplicação de forças musculares à malha da pele elasticamente conectada.

Waters [Waters1987] desenvolveu um modelo de face dinâmica que inclui dois tipos de músculos: músculos lineares que distendem (*pull*) e músculos *sphincter* que apertam (*squeeze*). Assim como o trabalho apresentado anteriormente, o trabalho de Waters usou um modelo mola-massa para a pele e para os músculos. Uma diferença existente é que os músculos de Waters têm propriedades direcionais (vetor) que são independentes da topologia da face específica. Cada músculo possui uma zona de influência, e a influência de um músculo particular é medida através de uma função da distância radial a partir do ponto do músculo em anexo.

1.3. AS EXPRESSÕES UNIVERSAIS

As expressões faciais humanas têm sido o assunto de um grande número de investigações na comunidade científica. Em particular, a questão da universalização das

expressões faciais através das diferentes culturas e as derivações de um pequeno número de expressões faciais principais têm consumido uma atenção considerável [Ekman1971].

Pesquisas na área de expressão facial levaram à conclusão de que existem seis categorias universais de expressões faciais: tristeza, raiva, alegria, medo, desgosto e surpresa, como ilustra a Figura 1-1 [Parke1996]. Dentro de cada uma dessas categorias pode existir uma variedade de "intensidades" das expressões faciais e algumas variações nos seus detalhes.

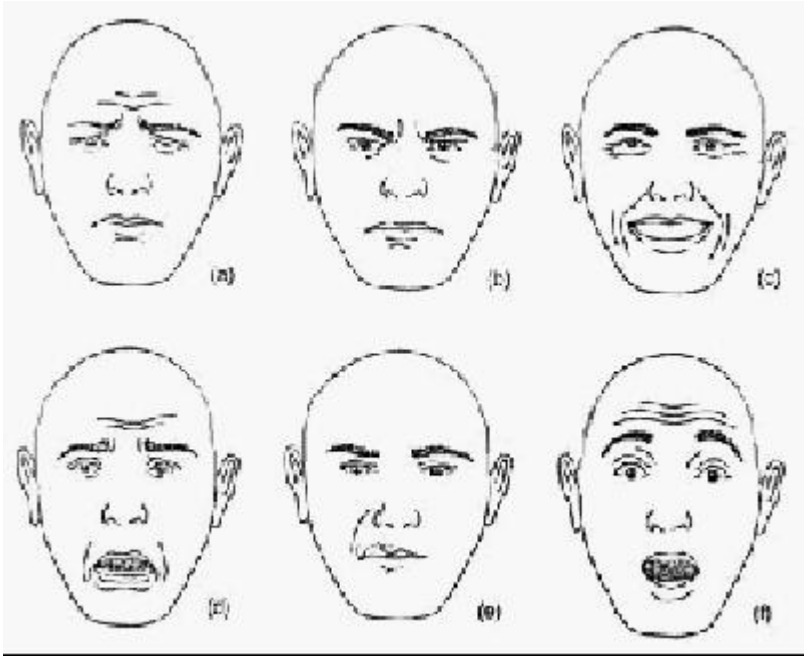


Figura 1-1: Expressões universais: (a) tristeza, (b) raiva, (c) alegria, (d) medo, (e) desgosto e (f) surpresa.

Cada uma dessas categorias possui traços característicos, principalmente nos componentes faciais olhos e boca, e nos locais onde as rugas se formam. Por exemplo, na categoria *alegria* (Figura 1-1(c)), as sobrancelhas estão relaxadas. As pálpebras superiores estão levemente abaixadas e as inferiores estão retas, sendo puxadas para cima pelas bochechas superiores. A boca fica extensa, com os cantos puxados para trás em direção às orelhas. Se a boca estiver fechada, os lábios ficam finos, conseqüentemente; se estiver aberta, os lábios superiores ficam retos, mostrando os dentes superiores, e os inferiores ficam retos no meio e angulares próximos aos cantos. As rugas para a alegria são "pés de galinha" nos cantos dos olhos, uma ruga em forma de sorriso se fixa abaixo das pálpebras inferiores, aparecem covinhas na face e no queixo e uma ruga profunda naso-labial do nariz até o queixo. As variações para a alegria são gargalhadas altas, gargalhadas, sorriso com a boca aberta, sorriso, sorriso melancólico, sorriso ávido, sorriso ingrato, sorriso malicioso, sorriso debochado, sorriso com olhos fechados, sorriso falso e gargalhada falsa.

Assim como a alegria, os traços característicos e as variações das demais categorias são descritos em [Parke1996] e em [Ekman1971].

1.4. PERSONAGEM REALISTA *VERSUS* PERSONAGEM DE CARTOON

Quando se pensa em desenvolver um sistema de animação facial uma das primeiras decisões, e talvez uma das mais importantes, é o tipo de personagem que se deseja animar. Basicamente, pode-se dividir o "tipo do personagem" em dois grupos: personagens realistas e personagens de estilo *cartoon* (ou caricatural).

O estilo realista busca ser o mais semelhante possível com a fisionomia da face humana. Justamente devido a sua semelhança com o ser humano, esse estilo de personagem é, na maioria dos casos, mas difícil de ser animado e exige uma maior precisão na animação facial produzida.

Em contraposição, o estilo caricatural destaca-se pela presença de fatores de distorção ou exagero da face específica, sendo normalmente utilizado na produção de personagens em desenhos animados. Como é possível "brincar" com o personagem caricatural, naturalmente a avaliação da animação é um pouco menos criteriosa, permitindo que o animador tenha mais leveza nos seus passos, "liberando" a sua criatividade.

1.5. ORGANIZAÇÃO DA MONOGRAFIA

O Capítulo 2 desta monografia apresenta o artigo "*Improvements on a simple muscle-based 3D face for realistic facial expressions*" [Bui2003]. Esse artigo tem sua publicação datada do ano de 2003 e é um artigo da Universidade de Twente, Holanda. Esse artigo descreve um modelo facial tridimensional simples com uma modelagem baseada em músculos físicos para definição e animação de expressões faciais realistas em tempo real.

O Capítulo 3 apresenta o artigo "*Geometry-Driven Photorealistic Facial Expression Synthesis*" [Zhang2003], publicado em 2003. Esse artigo apresenta um sistema de síntese de expressões faciais dirigida pela geometria (*geometry-driven*) e também apresenta o desenvolvimento de um editor de expressões onde o usuário modifica pontos característicos da face e o sistema interativamente gera a "nova" expressão facial.

O Capítulo 4 apresenta o artigo "*How Believable are Real Faces? Towards a perceptual basis for conversational animation*" [Cunningham2003], publicado em 2003. O artigo relata um experimento desenvolvido com pessoas a fim de determinar, de forma psicofisica, a ambigüidade e a credibilidade de expressões faciais.

O Capítulo 5 apresenta o artigo "*Unsupervised Learning for Speech Motion Editing*" [Cao2003], publicado em 2003. O artigo discute o problema de edição de movimento facial gravado, apresentando um sistema que faz uso da abordagem de aprendizado não-supervisionado através da técnica de análise de componentes independentes (ICA) para decompor o sinal da fala em componentes de conteúdo (a fala propriamente dita) e componentes de emoção.

O Capítulo 6 apresenta o artigo "*Learning Controls for Blend Shape Based Realistic Facial Animation*" [Joshi2003], publicado em 2003. Basicamente, o artigo propõe uma técnica automática de segmentação fisicamente motivada para aprender os controles e os parâmetros de uma animação diretamente do conjunto de *blend shapes*.

Essa técnica será utilizada, de forma eficiente, tanto para animação baseada na captura de movimento quanto para animação *keyframing*.

O Capítulo 7 apresenta o artigo "*An Example-Based Approach for Facial Expression Cloning"* [Pyun2003], publicado em 2003. Este artigo apresenta uma nova abordagem baseada em exemplos para clonagem de expressões faciais. A partir de um conjunto de exemplos de modelos-chave de uma determinada face origem, a idéia é conseguir uma clonagem das expressões faciais do modelo origem em um modelo destino (face destino) preservando as características faciais deste modelo alvo na animação final.

O Capítulo 8 apresenta o artigo *"Vision-based Control of 3D Facial Animation"* [Chai2003], publicado em 2003. Este artigo apresenta uma ferramenta desenvolvida em que um usuário qualquer faz uma expressão em frente a uma única câmera de vídeo e essa expressão é reproduzida por um avatar. O ponto interessante é que esse mesmo avatar é capaz de reproduzir as expressões de diferentes usuários em tempo real.

O Capítulo 9 é dedicado às conclusões desta monografia. Neste capítulo será apresentada uma comparação a respeitos dos artigos apresentados, suas correspondências e os pontos fortes apresentados na maioria deles. Um ponto importante deste capítulo também é uma motivação que é dada para o meu trabalho, em particular, por conseqüência da leitura e do aprendizado desses artigos: como eles podem contribuir e, principalmente, quais são os problemas em aberto em os artigos apontam e não foram trabalhados e porquê não foram. Por fim, o documento termina com uma lista das referência bibliográficas utilizadas no desenvolvimento desta monografia em "10.Referências Bibliográficas"

2. Artigo I: "Improvements on a Simple Muscle-Based 3D Face for Realistic Facial Expressions"

Esse artigo [Bui2003] apresenta um modelo de face tridimensional baseado em músculos físicos. Através do modelo facial é possível definir expressões faciais realistas como também animar a face de forma que ela modifique (translade) de uma expressão facial para outra.

2.1. INTRODUÇÃO

Quando se pensa em animação facial, é difícil dimensionar o intervalo e a variedade de faces e tipos de animação que podem ser desenvolvidos. Como visto na Seção 1.1, é possível definir uma face tanto através de sua geometria como através de imagens capturadas.

Esse artigo, em particular, define uma face tridimensional que é modelada através de sua forma geométrica. É importante destacar que o trabalho em questão está preocupado apenas com a forma e não com outras particularidades da face como características do personagem e até mesmo sua fala.

Dentre as possíveis técnicas de modelagem geométrica e animação para uma face (métodos de elementos finitos, imagens capturadas, interpolação, parametrização etc.), a face proposta neste artigo faz uso da técnica de um modelo físico baseado em músculos. É importante mencionar, como será visto mais adiante, que em alguns trechos faciais críticos houve necessidade de fazer uma modelagem de pseudo-músculo, sacrificando, conseqüentemente, um pouco do realismo facial.

O modelo definido pelos autores é uma extensão do modelo facial definido por Keith Waters em [Waters1987] (Seção 1.2.5), tendo sido adicionadas à face protuberâncias (*bulges*) e rugas (*wrinkles*). Além dessa adição das protuberâncias e das rugas, o modelo proposto também melhora a ação de múltiplos músculos e apresenta uma técnica para reduzir o tempo computacional de um modelo muscular.

A Figura 2-1 ilustra os músculos faciais definidos no modelo 3D de Keith Waters [Waters1987], enquanto que a Figura 2-2 ilustra três das possíveis expressões faciais que o modelo assume. Na Figura 2-2(a) tem-se a face em seu estado de neutralidade (emoção neutra), estando todos os músculos relaxados. Já na Figura 2-2(b) tem-se a expressão de surpresa. Como é possível verificar na imagem, as sobrancelhas estão curvadas e altas, as pálpebras abertas e as pupilas dilatadas. A emoção feliz, ilustrada na Figura 2-2(c), caracteriza-se por apresentar os lábios com seus cantos para trás levantados em direção oblíqua para o músculo zigomático.

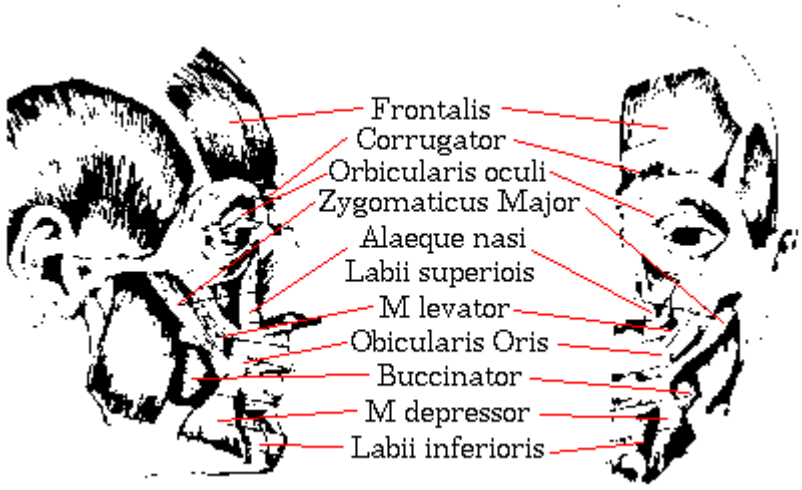
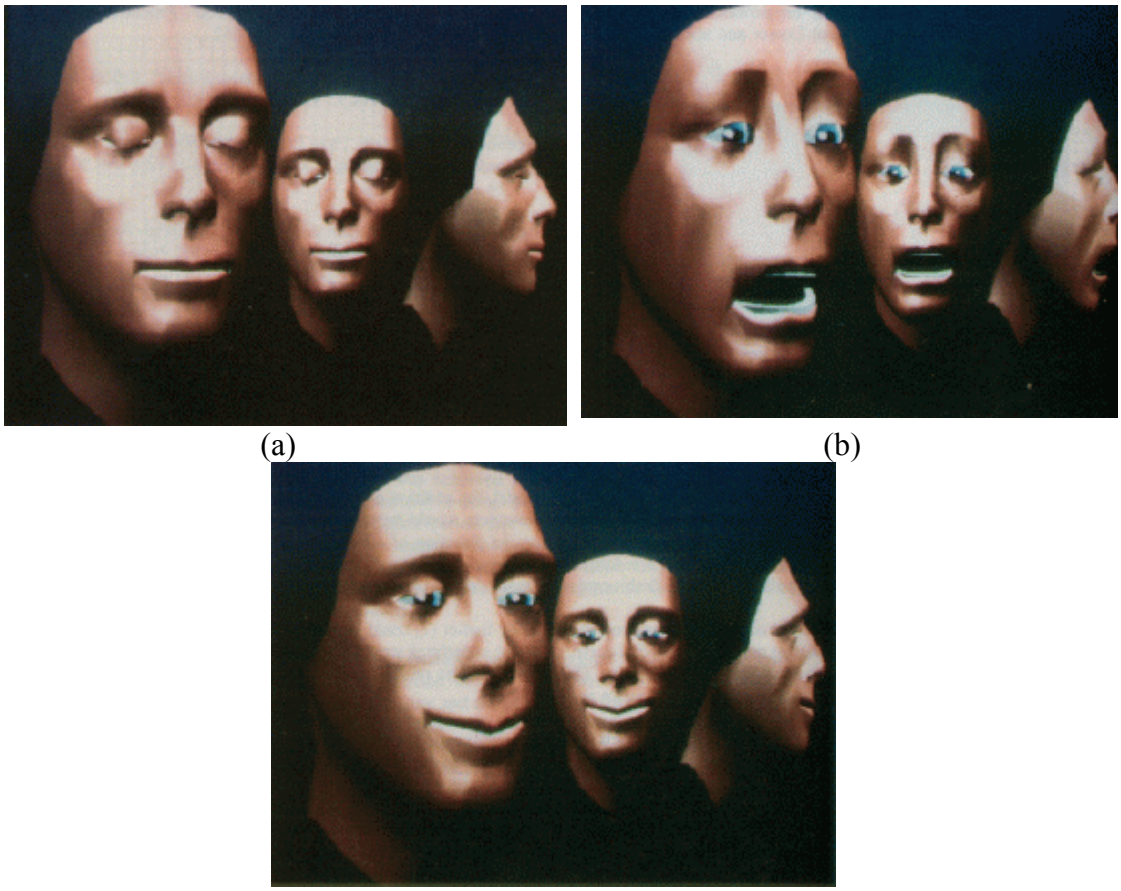


Figura 2-1: Modelo físico baseado em músculos do Keith Waters [Waters1987].



(a) (b)

(c)

Figura 2-2: Face nas suas expressões (a) neutra, (b) surpresa e (c) feliz.

2.2. O MODELO FACIAL

O modelo facial utilizado neste trabalho foi desenhado (renderizado) e animado utilizando *OpenGL* [Shreiner2003]. A estrutura poligonal permitiu uma deformação simples utilizando o modelo de músculos como também ela permite uma renderização rápida para computadores pessoais. Como os lábios são uma importante e difícil parte da face para animação, foi utilizada uma superfície *B-Spline* para a sua modelagem, permitindo assim um modelo mais realista e, conseqüentemente, podendo definir um movimento labial mais suave. Os lábios são importantes porque eles são manipulados tanto pelas expressões faciais quanto pelos visemas[1] no instante da fala [Lucena2002].

Os dados para a malha facial foram obtidos através de um *scanner* 3D. Uma vez que a face foi "escaneada", ela passou por duas outras etapas. A primeira etapa foi à redução do número de vértices e de polígonos de toda a face. A idéia foi simplificar a malha para gerar uma computação mais eficiente. Optou-se então por deixar mais vértices e polígonos nas partes mais expressivas, como nas regiões ao redor dos olhos, da boca e na testa. O modelo final da face 3D contém 2480 vértices e 4744 polígonos, como ilustra a Figura 2-3.

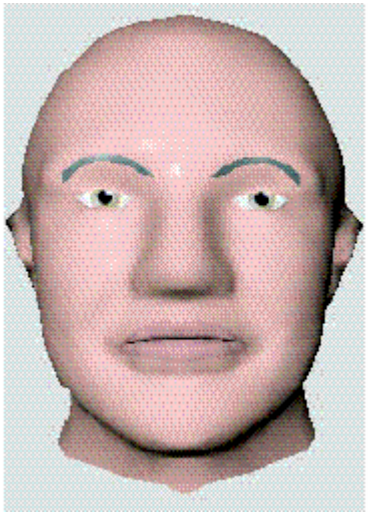


Figura 2-3: Modelo facial desenvolvido com 2480 vértices e 4744 polígonos. Face na sua expressão neutra.

A segunda fase consistiu em dividir a face em regiões, como ilustra a Figura 2-4. Baseado na distribuição dos músculos, a face foi dividida em 11 regiões: (1) face direita inferior (*right lower face*), (2) face direita central (*right middle face*), (3) pálpebra inferior direita (*right lower eyelid*), (4) pálpebra superior direita (*right upper eyelid*), (5) face direita superior (*right upper face*), (6) face esquerda inferior (*left lower face*), (7) face esquerda central (*left middle face*), (8) pálpebra inferior esquerda (*left lower eyelid*), (9) pálpebra superior esquerda (*left upper eyelid*), (10) face esquerda superior (*left upper face*) e (11) região não-animada (consistindo do restante da face). A Figura 2-4 ilustra o lado direito da face, com as suas respectivas regiões enumeradas.

[1] O "visema" pode ser conceituado como a representação visual de um fonema.

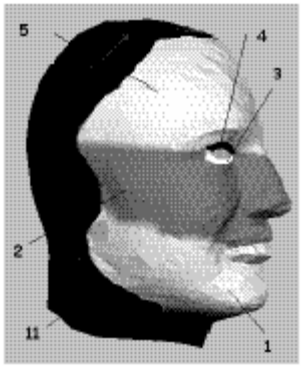


Figura 2-4: A região de divisão da face no seu lado direito.

Cada uma dessas 11 regiões em que a face foi dividida ainda foram divididas em sub-regiões. Por exemplo, tem-se como sub-região o *lábio superior*, que faz parte das regiões baixas da face (*right lower face* e *left lower face*), e as *sobrancelhas*, que pertencem às regiões superiores da face (*right upper face* e *left upper face*). Os vértices da malha da face são armazenados em um arquivo por regiões, como ilustra a Figura 2-5.

```
Lower face start vertex 0     Lower face end vertex 159
Middle face start vertex 160     Middle face end vertex 484 ...
```

Figura 2-5: Exemplo de arquivo que armazena os vértices da malha da face por regiões.

É possível afirmar que as duas regiões mais difíceis de animar são os lábios e os olhos [Bui2003]. Essas são as duas partes da face que contêm maior expressividade. Os lábios são difíceis principalmente quando a emoção da face está associada à fala, exigindo uma combinação dos visemas e da emoção propriamente dita. Os olhos são bastante expressivos, devendo levar em consideração tanto se as pálpebras estão abertas ou fechadas como também o formato e a posição que se encontram as sobrancelhas.

Como os lábios são a parte da face que possuem maior mobilidade, uma atenção especial foi dada a essa região no momento da modelagem facial. Conforme comentado anteriormente, o modelo labial deste artigo é uma superfície *B-Spline* com uma grade de controle de 24×6. Foi usada uma superfície *B-Spline* para aumentar a suavidade nos lábios após uma distorção causada pelo modelo muscular. Os lábios estão ilustrados na Figura 2-6.

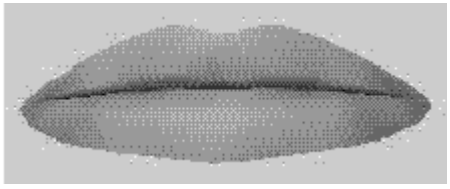


Figura 2-6: Os lábios do modelo facial definido.

É importante ressaltar que apesar de toda a preocupação no artigo com a mobilidade dos lábios em uma animação, o trabalho não considera a fala. Conseqüentemente, o artigo não se preocupa com a geração de visemas e suas combinações com as emoções.

Para os olhos, foi implementado o algoritmo de *tracking* descrito por Parke em sua tese de doutorado em 1974 [Parke1974]. O movimento dos olhos é independente dos movimentos dos músculos faciais.

2.3. OS MÚSCULOS QUE DIRIGEM A ANIMAÇÃO FACIAL

Os músculos que dirigem a animação facial neste trabalho são baseados no modelo do Keith Waters (Seção 1.2.5). Waters modelou três tipos de músculos: um vetor de músculos que é usado para a maioria dos músculos faciais, o músculo do tipo *sheet*, que é usado para os músculos frontais (*Frontalis*) e o músculo do tipo *sphincter* que é usado para o *Orbicularis Oris*.

O vetor de músculos é descrito na Figura 2-7. Esse modelo linear foi aplicado tanto para a maioria dos músculos faciais (o primeiro tipo descrito anteriormente), como se era esperado, como para os músculos frontais (o segundo tipo).

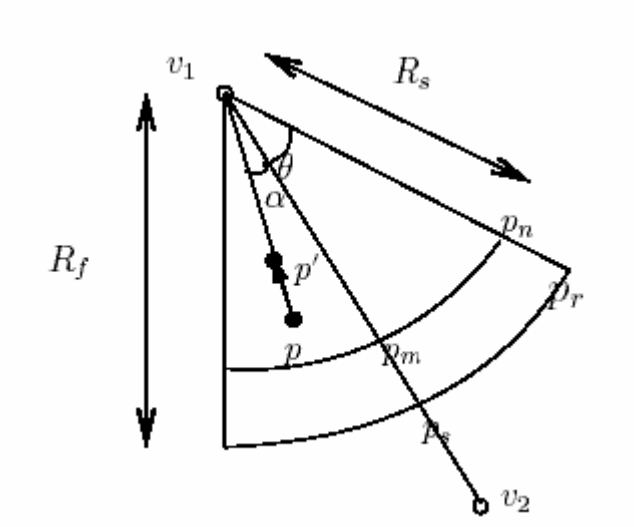


Figura 2-7: Modelo linear do músculo.

No caso do músculo *Frontails* (frontal), optou-se em usar o vetor de músculos ao invés do músculo do tipo *sheet* porque a testa não é completamente plana.

O músculo é modelado com um vetor de v_2 para v_1. R_s e R_f representam, respectivamente, o intervalo radial de início e fim. O novo vértice *p'* de um vértice arbitrário *p* localizado na malha dentro do segmento $v_1p_rp_s$, ao longo do vetor (p, v_1), é computado como:

$$p' = p + \cos(\alpha) kr \frac{pv_1}{\|pv_1\|}$$

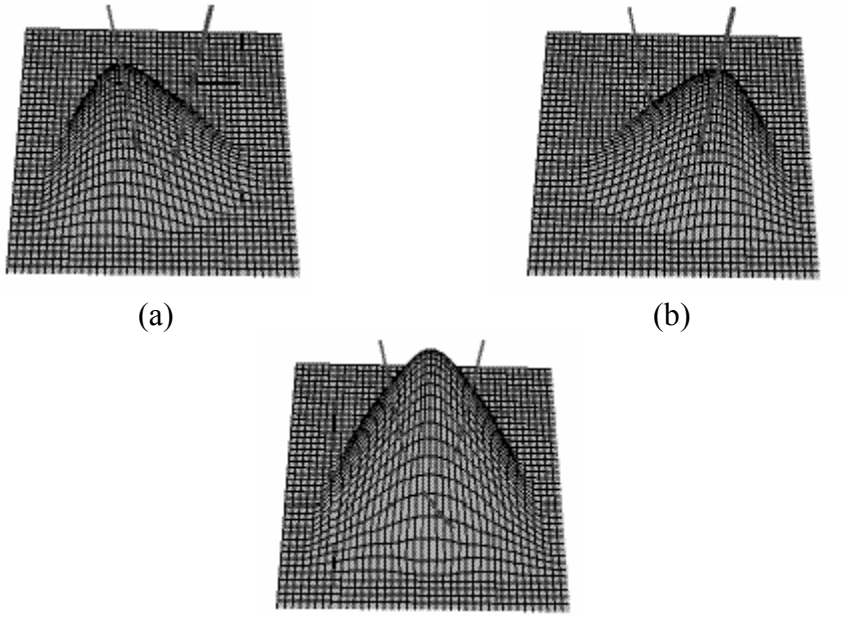
onde α é o ângulo entre o vetor (v_1, v_2) e (v_1, p), D é $\|v_1 - p\|$, k é uma constante fixa representando a elasticidade da pele e r é o parâmetro de deslocamento radial, dado por:

$$r = \cos\left(1 - \frac{D}{R_S}\right)$$ se p estiver dentro do setor ($v_1 p_n p_m$);

e

$$r = \cos\left(\frac{D - R_S}{R_f - R_S}\right)$$ se p estiver dentro do setor ($p_n p_r p_s p_m$).

Existe um problema associado com esse modelo, quando um vértice da malha está sobre a influência da ação de vários músculos. Dentre algumas soluções existentes e até mesmo já testadas, como a proposta por Keith Waters [Waters1987], é feita uma combinação das contrações musculares tentando simular o paralelismo dos músculos. Para um determinado vértice dentro de múltiplas zonas de influência, o artigo interativamente aplica unidades pequenas de níveis de contração ao vértice em questão, até que não haja mais contrações a serem aplicadas. Em cada vez, uma pequena unidade de contração δ_c é aplicada. Os deslocamentos dos vértices (causados pelos músculos que possuem o vértice em sua zona de influência) são somados e aplicados. Foi utilizado um $\delta_c = 0.2$ (o valor máximo para o nível de contração do músculo é de 1.0), o que, segundo os autores, produziu bons resultados. A Figura 2-8 mostra em (a) e (b) o efeito que ocorre em dois músculos da face e em (c) o resultado após aplicada a técnica do deslocamento.



(a) (b)

(c)

Figura 2-8: O efeito de um único músculo na malha em duas regiões em (a) e em (b), tendo em (c) o resultado da aplicação do deslocamento para correção.

Fisiologicamente, o *Orbicularis Oris* não é apenas o músculo *sphincter*, mas uma combinação de músculos que podem dirigir (mover) a boca em diferentes direções. A implementação utilizada neste artigo para o *Orbicularis Oris* é uma adaptação de [King2000].

Como mencionado anteriormente e ilustrado na Figura 2-6, os lábios foram modelados como uma superfície *B-Spline* com uma grade de controle de 24×6. No modelo do artigo, o músculo *Orbicularis Oris* atinge (atua) apenas na superfície do lábio e essa superfície é deformada pelo deslocamento dos pontos de controle.

O deslocamento de um ponto de controle p_i devido à contração do músculo *Orbicularis Oris* é descrito por:

$$p_i = o(\theta_i + e(p_i) + x_i)$$

onde *o* é o nível de contração do músculo *Orbicularis Oris*, θ é a rotação máxima devido ao *puckering* dos lábios e x_i é a extrusão máxima do músculo *Orbicularis Oris*. O $e_i(p)$ retorna o vetor de movimento por mover um ponto de controle *p* para um ponto da elipse criada pela contração do *Orbicularis Oris*. A Figura 2-9 ilustra a deformação dos lábios após a contração desse músculo.

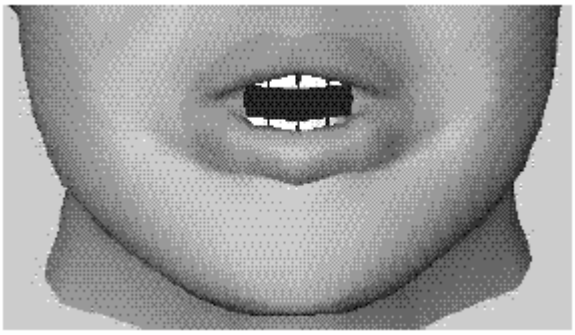


Figura 2-9: Deformação dos lábios devido à contração do músculo *Orbicularis Oris*.

Já o músculo *Orbicularis Oculi*, responsável pela movimentação dos olhos, possui duas partes: o *Pars Palpebralis*, que abre e fecha as pálpebras, e o *Pars Orbilatis*, que "aperta" (*squeeze*) os olhos. O artigo adaptou o algoritmo descrito em [King2000] para o abrir e o fechar das pálpebras.

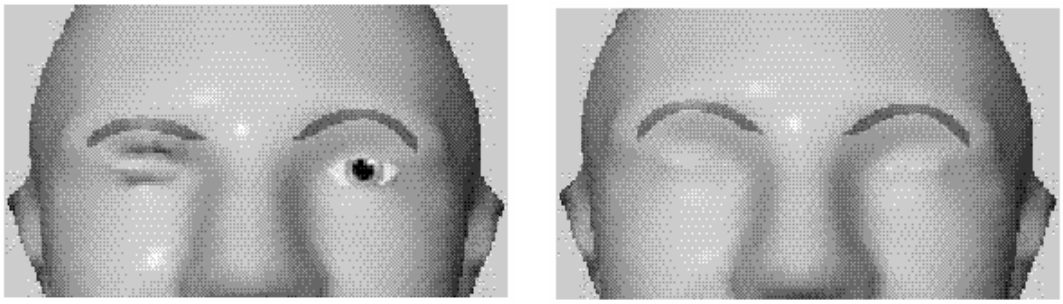
Uma pálpebra está aberta e fechada pela combinação da variação da técnica de mapeamento esférico com interpolação linear. A abertura dos olhos é implementada como um músculo *shincter* descrito em [Waters1987]. É importante notar que o fechar da pálpebra e o apertar de um olho ocorrem freqüentemente juntos.

A captura das relações existentes entre o abrir e o fechar de uma pálpebra e a abertura do olho foi feita pela modificação do nível de contração do *Pars Palpebralis* ($c_{closing}$) e o *Pars Orbitalis* ($c_{squeezing}$). A Figura 2-10 descreve o algoritmo utilizado para esse relacionamento.

$$\begin{aligned}&\text{if } c_{squeezing} > 0.5 c_{closing} \text{ then}\\ &\quad c_{closing} = max(1.0, 2 \times c_{squeezing})\\ &\text{else if try to close only one eye then}\\ &\quad c_{squeezing} = 0.5 c_{closing}\end{aligned}$$

Figura 2-10: Algoritmo de relacionamento entre o abrir e o fechar das pálpebras com a abertura dos olhos.

A Figura 2-11 ilustra em (a) o fechamento de um único olho e em (b) o fechamento de ambos os olhos.



(a) (b)

Figura 2-11: (a) O fechamento de um único olho e (b) o fechamento de ambos os olhos.

Uma outra região da face importante na animação e na formação das expressões faciais é a mandíbula (*jaw*). A mandíbula é aberta pela rotação dos vértices da parte inferior da face sobre o eixo pivô da mandíbula. O eixo de rotação é paralelo ao eixo X e passa através do ponto do pivô indicado da mandíbula. Os vértices localizados na região inferior da face são afetados pela rotação da mandíbula, tais como, o lábio inferior, os dentes superiores e os cantos da boca.

Com o objetivo de se ter uma boca oval com aparência natural, os vértices na parte inferior dos lábios são rotacionados em quantidades diferentes. Os vértices no meio do lábio inferior são rotacionados da mesma quantidade da rotação da mandíbula. A quantidade de rotação vai diminuindo na medida que vai se aproximando dos vértices dos cantos da boca. Os vértices dos cantos da boca são rotacionados por um terço da rotação da mandíbula.

O lábio superior também é afetado pela rotação da mandíbula. Os vértices do lábio superior são colocados para baixo ("*pulled down"*) com diferentes quantidades. A quantidade é zero para os vértices no meio do lábio superior e vai crescendo na medida que os vértices estão mais próximos dos cantos da boca.

Por fim, uma das contribuições deste artigo foi a adição de protuberâncias (*bulges*) e rugas (*wrinkles*) ao modelo facial apresentado em [Waters1987]. As protuberâncias e as rugas são criadas durante contrações dos músculos faciais.

Por questões de simplicidade, o artigo assumiu que os músculos são paralelos à pele facial e que as alturas das rugas para cada músculo são sempre as mesmas. Foram atribuídos valores predefinidos para a altura das rugas e o número de rugas (N_w) criadas pela contração de cada músculo. Esses valores são computados levando em consideração a preservação do volume e o modelo do esqueleto em questão.

A amplitude de uma ruga é calculada para todos os vértices que originalmente estão dentro da região $p_l p_k p_r v_3 p_t$, como ilustra a Figura 2-12. É importante mencionar que essa amplitude é calculada antes de se aplicar o deslocamento pela contração do músculo.

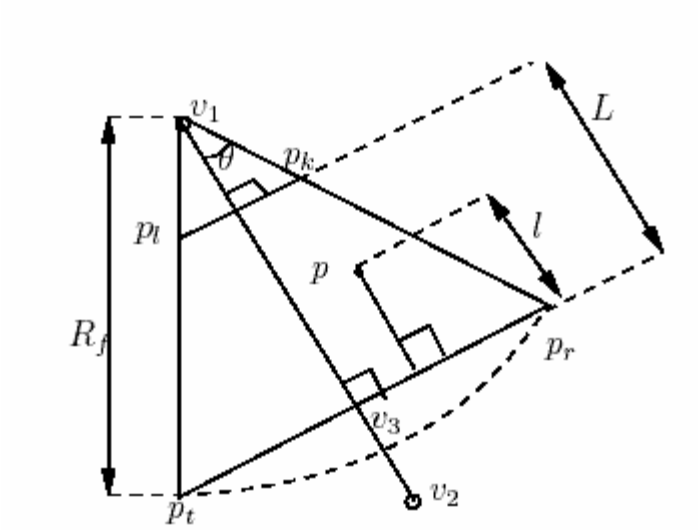


Figura 2-12: A zona que contém rugas devido à contração de um músculo linear.

A distância de p_l e de p_k para p_lp_r é dada por:

$$L = \frac{3}{4}|v_1v_3| = \frac{3}{4}R_f\cos(\theta)$$

A amplitude da ruga no vértice *p* é uma função da distância de *l* a partir de *p* para p_lp_r. O artigo utilizou uma série de parábolas para representar essa função, como ilustra a Figura 2-13, que é descrita pela seguinte equação:

$$amp(p) = f(l) = a\left(1 - \frac{(u-b)^2}{b^2}\right)$$

onde *a* é a altura das rugas, *b* é dado por $b = \frac{L}{2N_w}$ e *u* é dado por $u = l - \left\lfloor \frac{xN_w}{L} \right\rfloor \frac{L}{N_w}$.

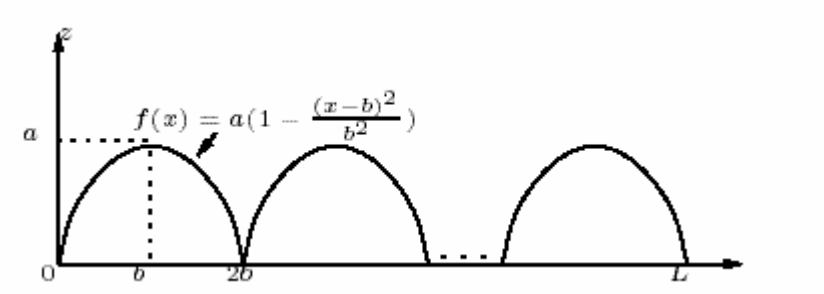


Figura 2-13: A função das rugas.

Para os vértices dentro da zona de influência de vários vértices, apenas a amplitude máxima da ruga causada para o músculo em questão é levada em consideração. As amplitudes das rugas são aplicadas depois dos vértices serem deslocados pela contração dos músculos.

Para tornar as rugas mais visíveis, o artigo previu o problema "*unrepresentative vertex normal*" (normal do vértice não representativa) devido à sombra interpolada para os vértices na parte de dentro das rugas, como ilustra a Figura 2-14. Esses vértices são os que possuem uma distância original para *ptpr* de *2b, 4b,..., L-2b*.

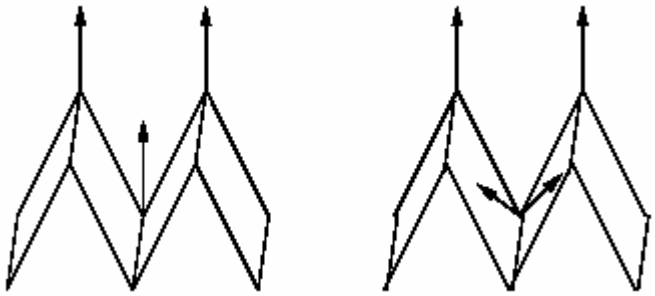
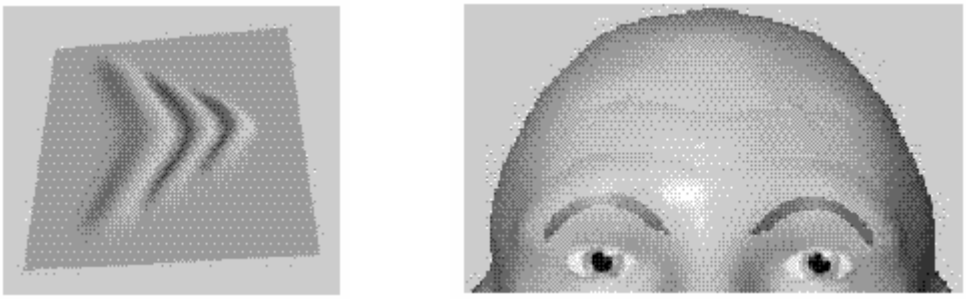


Figura 2-14: O problema do "*unrepresentative vertex normal*" e sua solução.

Ao invés de utilizar a normal do vértice não representativa, o artigo utilizou a normal dos polígonos triangulares que contém esses vértices. Um exemplo de ruga pode ser visualizado na Figura 2-15, onde em (a) tem-se a ruga propriamente dita e em (b) tem-se a aplicação da ruga na testa.



(a) (b)

Figura 2-15: Rugas devido às contrações musculares.

2.4. OS MELHORAMENTOS NA ANIMAÇÃO

A técnica para animação no modelo de músculos proposto por Waters seria, basicamente, como descrito na Figura 2-16.

```
for all vertices
    if the vertex is inside the muscle's zone of influence
      then calculate and apply the displacement of the vertex
```

Figura 2-16: Algoritmo inicial proposto para animação do modelo facial.

Esse é um bom algoritmo para um modelo de músculos lineares de uma face 3D com um número pequeno de polígonos. Como a complexidade desse algoritmo depende do número de vértices para verificar se cada vértice está ou não dentro da zona de influência do músculo, o mesmo se torna ineficiente para uma face com uma quantidade grande de vértices.

Como o modelo facial proposto neste artigo possui 2480 vértices e 4744 polígonos, foi necessário adicionar uma técnica de "*cut-off*" no algoritmo apresentado na Figura 2-16. A técnica introduzida está ilustrada a seguir:

```
for all vertices
    if the vertex is inside the region that the muscle has effect on
        if the vertex is inside the muscle's zone of influence
            then calculate and apply the displacement of the vertex
```

Figura 2-17: Extensão do algoritmo da Figura 2-16 para a face do artigo.

Como o modelo facial do artigo foi dividido em regiões e com o conhecimento das posições dos músculos faciais, ficou simples saber em que regiões um determinado músculo possui influência ou não. Cada músculo está associado com um *flag*, indicando que regiões da face o músculo tem ou não efeito.

Essa técnica primeiro elimina todos os vértices que estão fora da região que o músculo tem efeito, verificando apenas no arquivo descritor. Na medida que o número de vértices cresce, o tempo computacional de verificar tamanhos e ângulos irá cair fortemente. O melhoramento do algoritmo aplicado ao modelo facial deste artigo pode ser verificado na Figura 2-18.

	Animation speed
Before improv.	20.5fps
After muscle model improv.	30.5fps
After all improv.	35.0fps

Figura 2-18: Resultado do melhoramento da animação[2].

2.5. CONCLUSÕES DO ARTIGO

O artigo apresenta como principal contribuição o desenvolvimento de um modelo facial 3D simples que pode exibir expressões faciais realistas em tempo real. Alguns exemplos de expressões faciais são ilustrados na Figura 2-19, na Figura 2-20 e na Figura 2-21.

Dentre essas três expressões, a rotação da mandíbula pode ser bem ilustrada na emoção "feliz" (Figura 2-20). Já a face "surpresa" (Figura 2-19) ilustra bastante bem a

[2] Segundo o artigo, os testes foram feitos em uma máquina Pentium III com um processador de 800Mhz, utilizando 256MB de memória RAM e uma placa de vídeo NVidia GeForce3.

combinação dos músculos frontais gerando as rugas na testa. Por fim, na face "triste"(Figura 2-21) é possível verificar as protuberâncias e as rugas em algumas regiões.

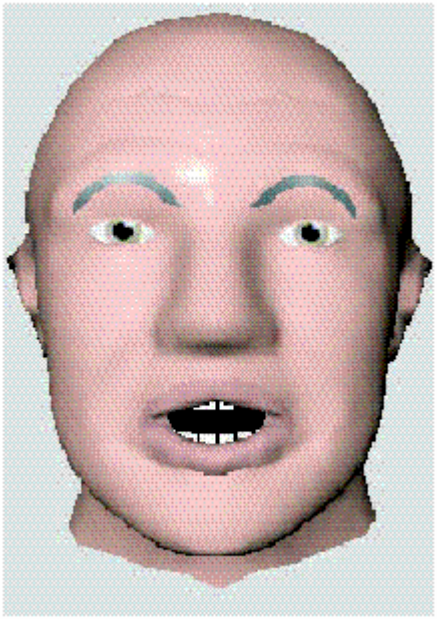


Figura 2-19: Modelo facial em sua expressão "surpresa".

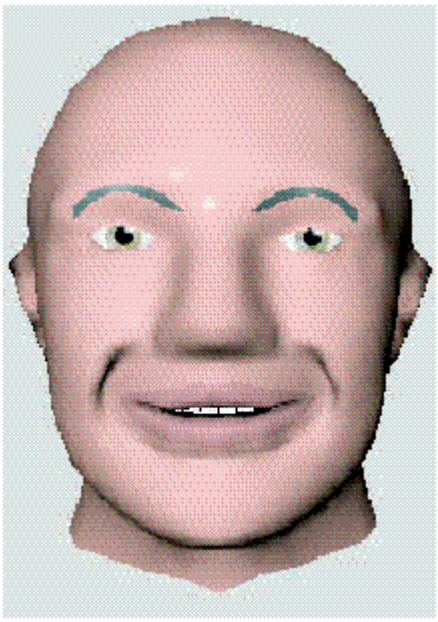


Figura 2-20: Modelo facial em sua expressão "feliz".

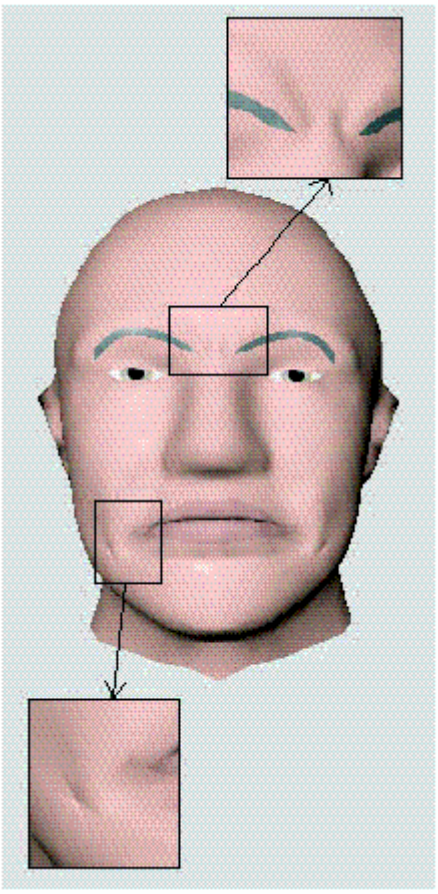


Figura 2-21: Modelo facial em sua expressão "triste".

Uma outra contribuição que o artigo destaca é que a partir da extensão do modelo facial de Keith Waters [Waters1987], foi possível gerar as protuberâncias e as rugas e ainda combinar várias ações musculares, conseguindo assim criar expressões faciais em tempo real. Mantendo ainda um certo nível de simplicidade, foi possível alcançar uma animação rápida em computadores pessoais.

O artigo ainda afirma que a abordagem desenvolvida é simples de aplicar em outras malhas, já que a representação muscular é independente da malha facial.

Como trabalhos futuros o artigo propõe os seguintes tópicos:

- Desenvolver um algoritmo de redução de malhas automático, que irá reduzir a necessidade de intervenção humana para geração de uma nova malha facial;
- Desenvolver um algoritmo para automaticamente mapear a divisão de regiões na nova malha facial; e;
- Testar como o mapeamento de textura irá afetar na qualidade das expressões faciais e na velocidade da animação.

2.6. CONCLUSÕES PESSOAIS

i. Quem são os autores?

The Duy Bui: Defendeu sua tese de doutorado recentemente na Universidade de Twente, Holanda (em julho de 2004). Basicamente, este artigo é um dos resultados apresentados em sua tese. O título da tese foi "Emotion and Facial Expressions in Creating Embodied Agents". Ele destaca como suas áreas de pesquisa: animação facial 3D, emoção e personalidade, expressões faciais e *embodied agents*.

Dirk Heyle: Foi co-orientador da tese de doutorado de The Duy Bui. É professor assistente do Departamento de Ciência da Computação da Universidade de Twente e trabalha no laboratório *Human Media Interaction*. Suas áreas de pesquisa são *multimodal interactions*, agentes inteligentes e *speech & language technology*.

Anton Nijholt: Um dos pesquisadores principais do Departamento de Ciência da Computação da Universidade de Twente. Suas principais áreas são *virtual reality & graphics*, *multimodal interactions*, agentes inteligentes, *speech & language technology* e *information engeneering*.

ii. O que o artigo resolve?

O artigo propõe um modelo facial tridimensional. A malha é definida utilizando um modelo muscular; apenas para os lábios é que se faz necessária uma abordagem de pseudo-músculos. A malha é composta por 2480 vértices que deram origem a 4744 polígonos. A partir da expressão de neutralidade e de aplicação de contrações musculares é possível, em tempo real, formar novas expressões faciais. O artigo afirma que suas expressões faciais são realistas e sua animação (translação de uma expressão facial para outra) ocorre de forma rápida, até mesmo se executada em um computador pessoal.

iii. Qual a abordagem utilizada?

A malha foi modelada através de OpenGL, tendo a biblioteca sido também utilizada para gerar a animação.

iv. Qual a classificação do artigo?

O artigo se enquadra num artigo de modelagem de expressões faciais.

v. Quais foram as ferramentas utilizadas na implementação?

O artigo não cita as ferramentas utilizadas no desenvolvimento do sistema, apenas menciona o uso da biblioteca OpenGL junto com um ferramental matemático para definição dos vetores de músculos e dos músculos para olhos, rugas e protuberâncias.

vi. Quais os problemas em aberto interessantes que o artigo aborda?

O artigo preocupa-se em modelar os lábios utilizando uma superfície *B-Spline*, que naturalmente vai favorecer uma maior suavidade no "desenho" dos lábios. O artigo também se preocupa com a interferência de um músculo

sobre outro, aplicando sucessivos níveis de contração até que não exista mais essa interferência.

vii. O que não foi feito neste artigo que é interessante? (Quais os problemas em aberto interessantes que o artigo não aborda?)

O artigo trata apenas da apresentação (modelagem) das expressões faciais no modelo 3D proposto. Apesar de sua preocupação com a modelagem dos lábios, visto que é a área facial de maior mobilidade, o artigo não associa o modelo a um sistema *talking head*. Como o próprio artigo menciona, ele não se preocupa com a formação dos visemas e, conseqüentemente, não aborda as limitações e problemas que ocorrem quando do casamento (sobreposição) dos visemas com as expressões faciais que movimentam os lábios de seu estado de neutralidade.

3. Artigo II: "Geometry-Driven Photorealistic Facial Expression Synthesis"

Este artigo [Zhang2003] apresenta um sistema de síntese de expressão facial dirigido por geometria (também chamado de animação baseada em performance – Seção 1.2.2). A partir das posições dos pontos característicos de uma expressão facial, o sistema é capaz de automaticamente sintetizar a imagem da expressão correspondente de forma fotorealista e com os detalhes da expressão bem aparentes. Uma outra aplicação do sistema é o editor de expressões onde o usuário pode mover os pontos característicos enquanto o sistema interativamente gera as expressões faciais com os detalhes de deformação da face.

3.1. INTRODUÇÃO

O artigo aponta a síntese de expressões faciais realistas como uma das áreas mais interessantes de pesquisa em computação gráfica como também uma das que possui problemas mais difíceis. O mapeamento de expressões (também chamado de *performance-driven animation*) tem sido um método popular para gerar animações faciais. Ele faz uso de movimentos de pontos característicos do usuário (*performer*) para dirigir os movimentos dos pontos característicos da face de uma pessoa diferente. Uma das falhas existentes nesse método é que o mesmo não produz detalhes nas expressões, tais como as rugas causadas pela deformação da pele.

Este artigo apresenta um sistema de síntese de expressões faciais dirigida pela geometria (*geometry-driven*). Dadas as posições dos pontos característicos de uma expressão facial, o sistema proposto automaticamente sintetiza a imagem da expressão correspondente com aparência fotorealista e natural, contendo os detalhes da expressão. Uma outra aplicação do sistema proposto é um editor de expressões, onde o usuário arrasta (modifica) os pontos característicos enquanto o sistema interativamente gera as expressões faciais com os detalhes de deformação da pele [Zhang2003].

3.2. TRABALHOS RELACIONADOS

Dentre os vários trabalhos relacionados apresentados no artigo, esta seção destina-se a apresentar os que foram configurados como de maior importância para a pesquisa que vem sendo desenvolvida neste trabalho. Um outro aspecto que foi utilizado para escolha dos trabalhos relacionados foi a maior proximidade com o próprio sistema que este artigo propõe a fim de realizar uma análise comparativa mais direta.

O trabalho de Noh e Neumann [Noh2001] é bastante interessante. O artigo apresenta uma técnica desenvolvida para clonagem de expressões onde são mapeados os movimentos geométricos da expressão de uma pessoa para uma outra pessoa diferente.

Uma abordagem eficiente para gerar expressões faciais fotorealistas com detalhes é a abordagem baseada em *morph*. Em particular, o trabalho de Pighin et al [Pighin1998] faz uso de combinações convexas das geometrias e das texturas de modelos de face exemplo para gerar expressões faciais fotorealistas. Eles também fornecem um conjunto

de ferramentas com interfaces fáceis de usar (*easy-to-use*) que permitem que o usuário interativamente projete expressões faciais. O sistema proposto por eles foi projetado para oferecer uma autoria *offline* e necessita que o usuário manualmente especifique os pesos das combinações para obter as expressões desejadas. O algoritmo de síntese proposto neste artigo [Zhang2003] diferencia-se do trabalho apresentado por Pighin et al [Pighin1998] no fato que o método deste artigo é completamente automático. Um outro aspecto, comparativamente importante, é que este artigo desenvolveu uma técnica para inferir os movimentos dos pontos característicos a partir de um subconjunto.

Um outro trabalho relacionado interessante é o de Liu et. al. [Liu2001]. Eles propuseram uma técnica chamada *expression ratio image*, onde os detalhes da expressão de uma pessoa são mapeados na face de outra pessoa diferente. Nesse método são dados os movimentos dos pontos característicos de uma expressão e é necessário também fornecer como entrada adicional a imagem de uma pessoa diferente com a mesma expressão. Em outras palavras, o método proposto por eles necessita de uma imagem de alguém para cada expressão diferente. Em contraste, o método apresentado neste artigo [Zhang2003] é capaz de gerar um número arbitrário de expressões a partir de um pequeno conjunto de imagens exemplo. Para as situações em que não há exemplos disponíveis para a face alvo, o método proposto em [Liu2001] é mais eficiente. Já para as situações em que são fornecidas as posições dos pontos característicos de uma expressão, mas nenhuma *expression ratio images* está disponível para essa geometria, o método proposto por este artigo é mais eficiente.

3.3. SÍNTESE DE EXPRESSÃO DIRIGIDA POR GEOMETRIA

O problema em questão nesta seção é que a partir das posições dos pontos característicos de uma expressão facial objetiva-se computar a imagem da expressão facial correspondente. Uma possível solução para essa questão seria utilizar algum mecanismo de simulação física. O problema apresentado na abordagem de simulação física para esse cenário se deve ao fato de se tornar difícil modelar deformações detalhadas da pele, tais como rugas, e também é difícil renderizar um modelo que tenha aparência fotorealística.

Dado um conjunto de expressões, Pighin et. al. [Pighin1998] demonstrou que é possível gerar expressões faciais fotorealistas através de uma combinação convexa. Seja $E_i = (G_i, I_i), i = 0,...,m$ expressões exemplo, onde G_i representa uma geometria e I_i representa uma imagem de textura. O artigo assume que todas as imagens de textura estão alinhadas por *pixel*. Seja $H(E_0, E_1,..., E_m)$ o conjunto de todas as possíveis combinações convexas desses exemplos. Então

$$H(E_0, E_1,..., E_m) = \left\{ \left(\sum_{i=0}^{m} c_i G_i, \sum_{i=0}^{m} c_i I_i \right) \middle| \sum_{i=0}^{m} c_i = 0, c_i \geq 0, i = 0,...,m \right\}$$

Pighin et. al. [Pighin1998] também desenvolveram um conjunto de ferramentas onde o usuário pode usá-las para interativamente especificar os coeficientes c_i para gerar as expressões desejadas.

Note que cada expressão no espaço $H(E_0, E_1, ..., E_m)$ possui um componente geométrico dado por $G = \sum_{i=0}^{m} c_i G_i$ e um componente de textura dado por $I = \sum_{i=0}^{m} c_i I_i$. Como o componente geométrico é bem mais fácil de ser obtido, o artigo propôs usar esse componente para inferir o componente de textura.

Dado o componente geométrico G, é possível projetar G para um fecho convexo expandido $G_0, G_1, ..., G_m$ e depois utilizar os coeficientes resultantes para compor as imagens exemplo e obter a imagem da textura desejada.

Um problema existente nessa abordagem é que o espaço de $H(E_0, E_1, ..., E_m)$ é muito limitado. Uma pessoa pode ter expressões de rugas em diferentes regiões da face, gerando muitas opções de combinação. Para resolver esse problema, o artigo propôs subdividir a face em um número de sub-regiões. Para cada sub-região é usada a geometria que está associada com essa sub-região para computar a sua imagem de textura. Depois essas imagens são combinadas (*seamlessly*) para produzir a imagem final da expressão.

Uma alternativa em potencial dessa combinação convexa é simplesmente utilizar um espaço linear, sem adicionar restrições aos coeficientes $c_i's$. O problema dessa alternativa é que os coeficientes resultantes da aproximação das geometrias por um espaço linear podem ser coeficientes negativos, como também serem maior que 1. Isto causaria falhas nas imagens compostas.

O artigo então apresenta o algoritmo desenvolvido para casos 2D (bidimensionais), onde a geometria de uma expressão é composta por pontos característicos projetados em uma imagem plana. Em seguida, o artigo apresenta a extensão do algoritmo para o caso 3D (tridimensional).

3.4. VISÃO GERAL DO SISTEMA

A Figura 3-1 apresenta uma visão geral do sistema. O sistema apresentado consiste, basicamente, de uma unidade *offline* (em tempo de autoria/pré-compilação) e uma unidade em tempo de execução. As imagens exemplo são processadas em tempo *offline* uma única vez. Em tempo de execução, o sistema captura as posições dos pontos característicos, fornecidos como entrada para a nova expressão, e produz a imagem da expressão final.

As seções seguintes destinam-se a apresentar cada uma dessas etapas em maiores detalhes.

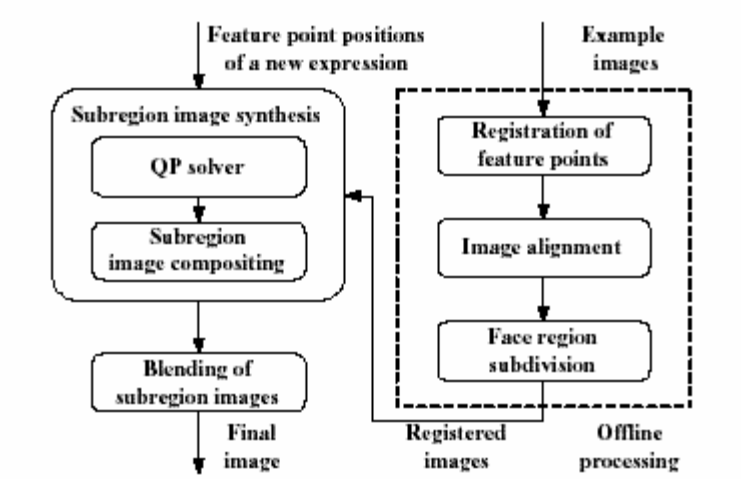


Figura 3-1: Visão geral do sistema de síntese de expressões *geometry-driven*.

3.5. PROCESSAMENTO *OFFLINE* DAS IMAGENS-EXEMPLO

A Figura 3-2 ilustra os pontos característicos que são utilizados no sistema proposto por este artigo [Zhang2003]. Na parte inferior esquerda da figura é possível ver os pontos característicos para os dentes, necessários quando a boca está aberta. No total foram 134 pontos característicos.

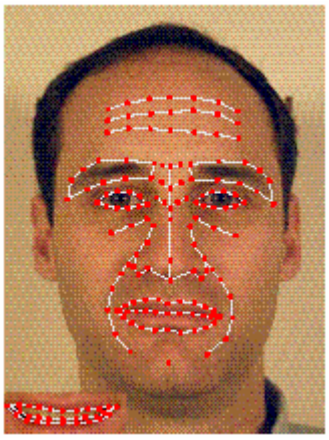


Figura 3-2: Pontos característicos da face.

Dada uma imagem de uma face, o sistema é capaz de computar automaticamente as características da face [Li2001]. Como o número de imagens exemplo é muito pequeno no sistema proposto em [Zhang2003] (10 a 15 imagens), os autores escolheram marcar manualmente pontos característicos das imagens exemplo.

Depois de obtida as marcas dos pontos característicos, todas as imagens exemplo são alinhadas com uma "imagem padrão", ilustrada na Figura 3-3. A razão de se criar uma "imagem padrão" deve-se ao fato de que era necessário ter a boca aberta para assim se ter uma textura para os dentes. O alinhamento é feito através de uma triangulação baseada no *warping* da imagem.

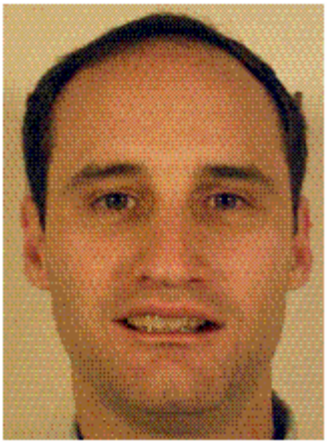


Figura 3-3: Imagem padrão.

A Figura 3-4 ilustra a divisão da face em 14 sub-regiões. Os autores procuraram ter sub-regiões pequenas até o ponto de evitar que rugas das expressões cruzassem os seus limites.

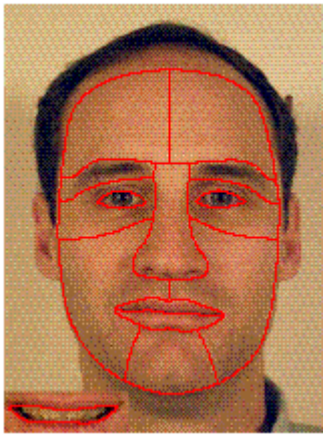


Figura 3-4: Região da face subdividida.

Uma vez que todas as imagens exemplo já foram alinhadas, o passo seguinte constituiu em subdividir a imagem padrão. Os autores criaram uma máscara para armazenar as informações de subdivisão, onde cada *pixel* guardava a informação de que sub-região ele pertencia no seu canal de cor.

3.6. SÍNTESE DA EXPRESSÃO DA SUB-REGIÃO

Seja n o número de pontos característicos, para cada expressão exemplo E_i, foi usado um G_i para denotar o vetor bidimensional $2n$ que consiste de todas as posições dos pontos característicos. Seja G as posições dos pontos característicos da nova expressão, para cada sub-região R, foi utilizado G_i^R para denotar os pontos característicos de E_i, os quais estão dentro ou no limite de R. Similarmente, o artigo usou G^R para denotar os pontos característicos de G associados a R.

Dado G^R, deseja-se projetar o fecho convexo de $G_0^R,...,G_m^R$. Em outras palavras, deseja-se encontrar o ponto mais próximo do fecho convexo. Isto pode ser formulado como um problema de otimização:

Minimizar: $\left(G^R - \sum_{i=0}^{m} c_i G_i^R\right)^T \left(G^R - \sum_{i=0}^{m} c_i G_i^R\right)$

Supondo: $\sum_{i=0}^{m} c_i = 1$

$c_i \geq 0, i = 0,1,...m$

Denotando

$$G = \left(G_0^R, G_1^R, ..., G_m^R\right)$$

e

$$C = \left(c_0, c_1, ..., c_m\right)$$

Então a função objetiva é dada por $C^{\mathrm{T}}\ G^{\mathrm{T}}\ G\ C - 2G^{R^T}\ G\ C + G^{R^T} G^R$.

Isso é uma formulação de programação quadrática onde a função objetiva é uma forma semi-definida quadrática positiva e suas restrições são lineares. Como os G_i^R's são, em geral, linearmente independentes, a função objetiva é, em geral, definida positivamente.

Existem vários caminhos para se resolver o problema de programação quadrática. O artigo [Zhang2003] aborda alguns deles, mas por questões de resumo e clareza vou apresentar apenas o que, na prática, foi utilizado no artigo.

Foi utilizado o método de *interior point* (ponto interior) [Ye1997]. Esse método trabalha pela interação no interior do domínio que é restringido por restrições de desigualdade (*inequality*). A cada interação, ele faz uso de uma extensão do método de Newton para encontrar o próximo ponto característico que é mais próximo do ótimo. Comparado com as abordagens tradicionais, o método do ponto interior apresenta uma taxa de convergência mais rápida tanto teoricamente quanto na prática, além de ser numericamente estável. Apesar dessa abordagem, em geral, não produzir a solução ótima, sua solução é, normalmente, bastante próxima da ótima. Em relação à aplicação desse método em [Zhang2003], os autores julgaram que o método atendia aos requisitos.

Depois de obter os coeficientes c_i's, o passo seguinte consiste em computar a imagem da sub-região I^R pela composição das imagens exemplo:

$$I^R = \sum_{i=0}^{m} c_i I_i^R$$

É importante relembrar que as imagens exemplo já foram alinhadas. Portanto, esse passo resume-se apenas à combinação do canal de cor dos *pixels*.

3.7. COMBINANDO NOS LIMITES DAS SUB-REGIÕES

Para evitar descontinuidade na imagem, os autores aplicaram uma combinação *fade-in-fade-out* nos limites das sub-regiões. Na implementação do artigo foi utilizado

um mapa de pesos para facilitar essa combinação (*blending*). A Figura 3-5 ilustra o mapa de pesos, o qual está alinhado com a "imagem padrão" (Figura 3-3).

Figura 3-5: Mapa de pesos para aplicar a combinação nos limites das sub-regiões.

As curvas grossas vermelhas e pretas são as regiões combinadas ao longo dos limites da curva. A intensidade do canal R (vermelho) armazena o peso da combinação. O artigo utilizou os canais G (verde) e B (azul) para armazenar os índices das duas sub-regiões vizinhas, respectivamente. Dado um *pixel* na região de combinação, e sendo r o valor do canal R e i_1 e i_2 os índices das duas sub-regiões, a intensidade da combinação é dada por:

$$I = \frac{r}{255} * I^{i_1} + \left(1 - \frac{r}{255}\right) * I^{i_2}$$

É importante observar que os autores não executaram o *blending* em alguns limites, onde, por exemplo, existe uma cor natural de descontinuidade, como nos olhos e nos lábios.

Após a combinação, uma imagem é obtida, estando esta já alinhada com a "imagem padrão". O passo seguinte é fazer o *warp* dessa imagem de tal forma que as posições dos seus pontos característicos coincidam com as posições dos pontos característicos fornecidos como entrada. Uma vez feito isso, obtém-se a imagem final da expressão.

3.8. DENTES

Como a região dos dentes é quase que ortogonal às outras regiões da face, o artigo optou por utilizar um conjunto separado de exemplos para a região dos dentes. No sistema atual, apenas um conjunto pequeno de exemplos foi utilizado para os dentes. Isso se deve ao fato de que o artigo não está focado em animações com fala, onde existem muitas variações nos formatos da boca.

3.9. SÍNTESE DA EXPRESSÃO EM 3D

Esta subseção dedica-se a apresentar a metodologia que foi utilizada, para estender o algoritmo apresentado para o caso 3D (tridimensional), onde os pontos característicos são posições 3D, e as expressões são malhas 3D, com ou sem mapeamento de textura. Para computar os coeficientes de combinação das sub-regiões, foi utilizada a equação

$G = \left(G_0^R, G_1^R, ..., G_m^R\right)$ da mesma forma que a anterior, exceto pelo fato que G e G_i são vetores de dimensão *3n*.

Foi utilizado o mesmo método do ponto interior (*interior point*) para resolver o problema de programação quadrática. Para a composição da malha de sub-região e para a combinação nos limites das sub-regiões também foi utilizada uma abordagem semelhante ao caso 2D, exceto pelo fato que estavam sendo combinadas posições de vértices 3D ao invés de imagens.

3.10. Inferindo os Movimentos dos Pontos Característicos a partir de um Subconjunto

Na prática é difícil obter todos os pontos característicos necessários, ilustrados na Figura 3-2. A maioria dos algoritmos de *track* (rastreamento) de faces rastreia (captura) um número limitado de pontos característicos ao longo das sobrancelhas, olhos, bocas e nariz. No mapeamento melhorado de expressões, que é discutido na subseção 3.11, os autores conseguem extrair apenas 40 pontos característicos do *performer*. Já na aplicação de edição de expressão, que será apresentada na subseção 3.12, cada vez que o usuário move um ponto característico o sistema precisa deduzir qual será o movimento mais próximo do correto para os pontos característicos restantes.

Essa subseção destina-se a descrever como inferir os movimentos para todos os pontos característicos de um mesmo subconjunto. Os autores assumiram uma abordagem baseada em exemplos. Com o objetivo de ter um controle com granularidade fina, que é particularmente importante se apenas os movimentos de um número muito pequeno de pontos característicos estiverem disponíveis, tal como na edição de uma expressão, os pontos característicos da face são divididos em hierarquias e é aplicada uma análise de componentes principais hierárquicas nas expressões exemplo.

No nível de hierarquia 0 tem-se um único conjunto de pontos característicos que controla o movimento global da face inteira. Na hierarquia 1 tem-se quatro conjuntos de pontos característicos, cada um controlando o movimento local das regiões características da face (região do olho esquerdo, região do olho direito, região do nariz e região da boca). Na hierarquia 2, cada conjunto de pontos característicos controla detalhes de regiões da face, tais como, formato das pálpebras, formato da linha dos lábios etc. Existem 16 conjuntos de pontos característicos no nível de hierarquia 2 e eles são usados como uma ponte entre os movimentos locais e globais da face de tal forma que é possível propagar os movimentos dos vértices de uma hierarquia para outra.

Para cada conjunto de pontos característicos, o sistema calcula o deslocamento de todos os vértices pertencentes ao conjunto característico para cada expressão exemplo. Depois é aplicada uma análise de componentes principais nos vetores de deslocamento do vértice correspondente às expressões e gerado o vetor de menor espaço dimensional.

3.10.1. Propagação do Movimento

Esta subseção destina-se a descrever como utilizar o resultado da análise de componentes principais hierárquicos para propagar nos movimentos dos vértices. O objetivo é, a partir do movimento de um subconjunto de pontos característicos, poder inferir o movimento mais razoável para o restante dos pontos característicos. A idéia

básica é aprender, a partir dos exemplos, como o restante dos pontos característicos deve se mover quando um subconjunto de vértices se move.

Seja $v_1, v_2, ..., v_n$ todos os pontos característicos de uma face. Seja δV o vetor deslocamento de todos os pontos característicos. Para um dado δV e um conjunto de pontos característicos F (o conjunto de índices dos pontos característicos pertence a esse conjunto de pontos característicos), foi utilizado $\delta V(F)$ para denotar o sub-vetor desses vértices que pertencem a F. Seja $Proj(\delta V, F)$ a projeção de $\delta V(F)$ no subespaço expandido pelas componentes principais correspondentes a F. Em outras palavras, $Proj(\delta V, F)$ é a melhor aproximação de $\delta V(F)$ no subespaço da expressão. Dados δV e $Proj(\delta V, F)$, é possível afirmar que c é atualizado por $Proj(\delta V, F)$ se para cada vértice que pertença a F for feita a substituição do deslocamento em $\delta V(F)$ pelo seu valor correspondente em $Proj(\delta V, F)$.

A Figura 3-6 descreve o algoritmo de propagação do movimento. Para facilitar o entendimento, o artigo descreve como inferir os movimentos de todos os pontos característicos a partir do movimento de um único vértice. Assuma que v_i é o vértice de movimento e foi obtido o vetor δV, onde δv_i é igual ao deslocamento para o vértice v_i, enquanto que o deslocamento dos restantes dos vértices é 0. Para propagar o movimento do vértice, o passo inicial é encontrar o conjunto de pontos característicos F^* que é a menor hierarquia entre todos os conjuntos de pontos característicos que contém v_i.

O algoritmo procede da seguinte forma: para cada conjunto de pontos característicos F, é utilizado o *flag* `hasBeenProcessed`(F) para denotar se F foi processado ou não. Inicialmente esse *flag* é definido como `falso` para todo F.

$MotionPropagation(F^*)$

Begin

Set h to be the hierarchy of F^*.

If $hasBeenProcessed(F^*)$ is true, return.

Compute $Proj(\delta V, F^*)$.

Update δV with $Proj(\delta V, F^*)$.

Set $hasBeenProcessed(F^*)$ to be true.

For each feature set F belonging to hierarchy $h-1$ such that $F \cap F^* \neq \emptyset$

$MotionPropagation(F)$

For each feature set F belonging to hierarchy h+1 such that $F \cap F^* \neq \emptyset$

$MotionPropagation(F)$

End

Figura 3-6: Algoritmo de propagação do movimento.

De forma similar, é possível inferir os movimentos de todos os pontos característicos de um subconjunto. Assumindo que o subconjunto de pontos característicos $v_{i_1}, v_{i_2}, ..., v_{i_k}$ tem movimento, define-se o vetor δV de tal forma que δv_{i_j} é igual ao vetor deslocamento para os vértices v_{i_j}, onde $j = 1,...,k$. Para cada vértice v_{i_j}, encontra-se o conjunto de pontos característicos F^j tal que seja a menor hierarquia entre todos os conjuntos de pontos característicos que contêm v_{i_j} e execute o `MotionPropagation`(F^j) (note que agora δV contém o deslocamento para todos v_{i_j}, $j = 1,...,k$).

3.11. MELHORANDO O MAPEAMENTO DA EXPRESSÃO

A técnica de mapeamento de expressões (também chamada de animação baseada em performance – *performance-driven animation*) é uma técnica simples e muito utilizada em animação facial. Ela trabalha através da computação do vetor diferença das posições de pontos característicos entre a face no seu estado neutro (expressão natural) e a expressão da face do *performer*, e depois adiciona o vetor diferença à geometria da face do novo personagem. Uma das desvantagens na expressão facial resultante é que ela pode não parecer convincente no que diz respeito aos seus detalhes.

A técnica proposta neste artigo fornece uma solução para esse problema. A idéia consiste em obter imagens exemplo para o novo personagem. As imagens exemplo podem ser obtidas no processamento offline, através da captura ou da projeção de um artista. Em tempo de execução, primeiro usa-se o vetor de diferença geométrica para obter a geometria desejada para o novo personagem como no sistema tradicional de mapeamento de expressão. Devido à dificuldade do *tracking* da face, o número de pontos característicos disponíveis é bem menor do que o número necessário pelo sistema de síntese. Para resolver essa questão, é utilizada a técnica apresentada na subseção 3.3 para inferir os movimentos para todos os pontos característicos utilizados pelo sistema de síntese. Depois é aplicada a técnica de gerar a imagem de textura baseada na geometria. Os resultados finais são bem mais convincentes e têm expressões faciais bem mais realistas.

3.12. EDITANDO A EXPRESSÃO

Outra aplicação interessante desenvolvida neste artigo é um editor de expressões interativo. A idéia é que o usuário possa mover pontos característicos da face e o sistema, interativamente, exiba a imagem resultante com os detalhes da expressão. A Figura 3-7 é uma tela capturada da interface do editor de expressões, onde os pontos em vermelho representam os pontos característicos que o usuário pode clicar e arrastar.

O primeiro estágio do sistema é a geração da geometria. Quando o usuário arrasta um ponto característico, o gerador de geometria infere as "mais prováveis" posições para todos os outros pontos característicos usando o algoritmo descrito na subseção 3.3. Por exemplo, se o usuário arrasta um ponto característico do topo do nariz, toda a região do nariz irá mover.

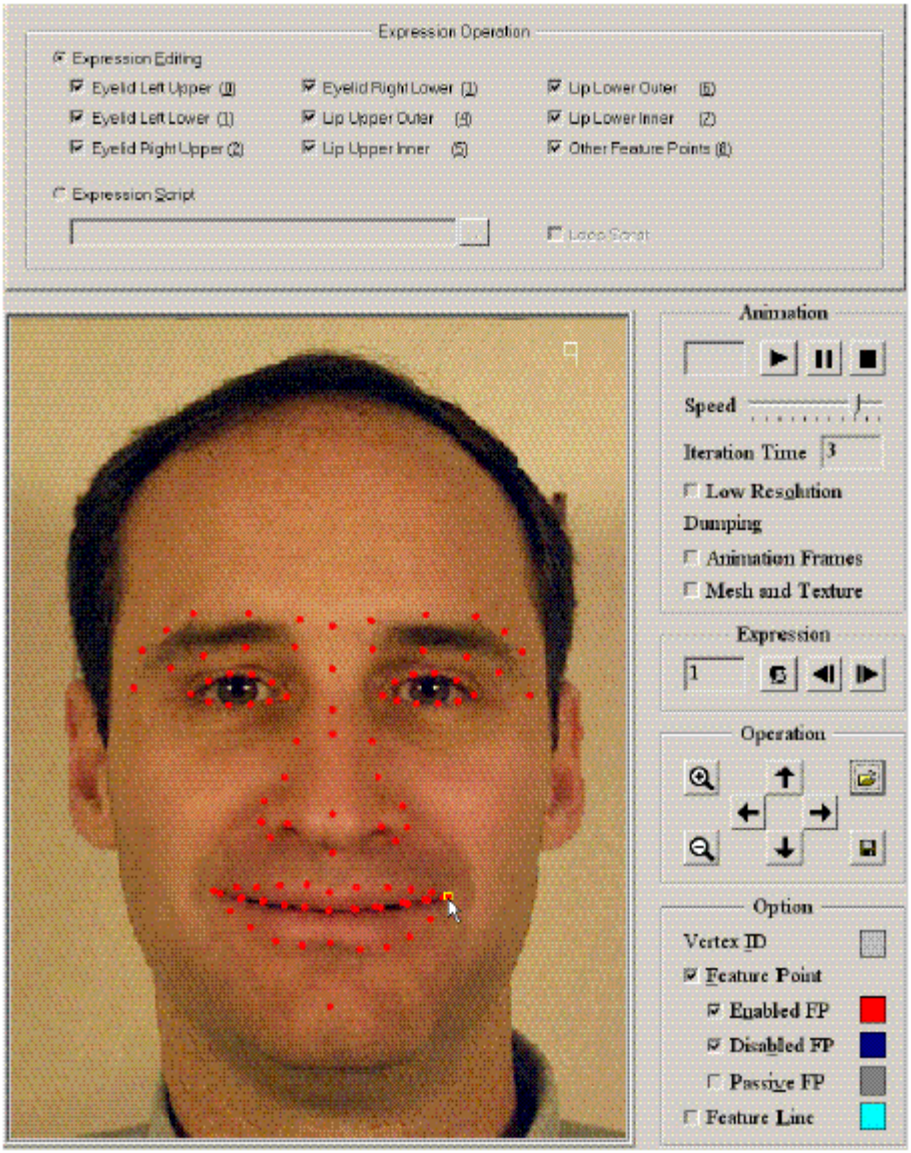


Figura 3-7: Interface do editor de expressões. Os pontos em vermelho são os pontos característicos que o usuário pode clicar e arrastar.

Tipicamente foram utilizadas 30 a 40 imagens-exemplo para a inferência do ponto característico tanto na aplicação de edição de expressão quanto na aplicação de mapeamento de expressão.

3.13. RESULTADOS

O artigo mostra os resultados obtidos para duas faces: uma masculina e uma feminina. Para cada pessoa foram capturadas cerca de 30 a 40 imagens com todos os tipos de expressão que eles podiam fazer.

Depois as imagens exemplo foram selecionadas e as imagens restantes foram utilizadas para validar o sistema. Serão apresentados os resultados para o grupo de imagens masculinas, para os demais exemplos consultar o próprio artigo [Zhang2003].

A Figura 3-8 ilustra as imagens exemplo para o homem, enquanto a Figura 3-9 ilustra os exemplos para os dentes.

Figura 3-8: Imagens-exemplo para o homem.

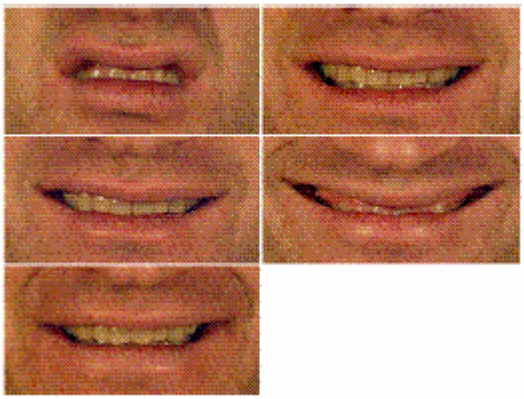


Figura 3-9: Imagens-exemplo para os dentes do homem.

A Figura 3-10 compara, lado a lado, as imagens. Onde do lado esquerdo têm-se as imagens originais e do lado direito têm-se as imagens resultantes da síntese.

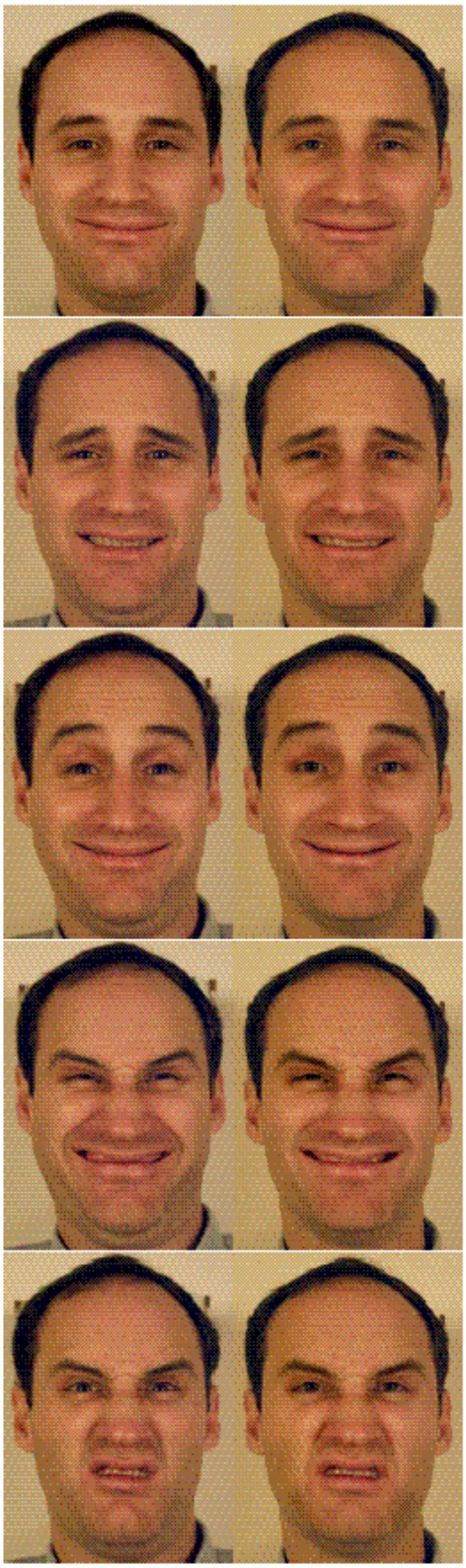


Figura 3-10: Comparação lado-a-lado das imagens verdadeiras (coluna da esquerda) com os resultados sintetizados (coluna da direita).

A Figura 3-11 ilustra o melhoramento do mapeamento de expressão, ilustrando alguns dos resultados de mapear as expressões femininas no homem. As expressões femininas são dados reais e o resultado da direita (expressões masculinas) são os resultados do mapeamento.

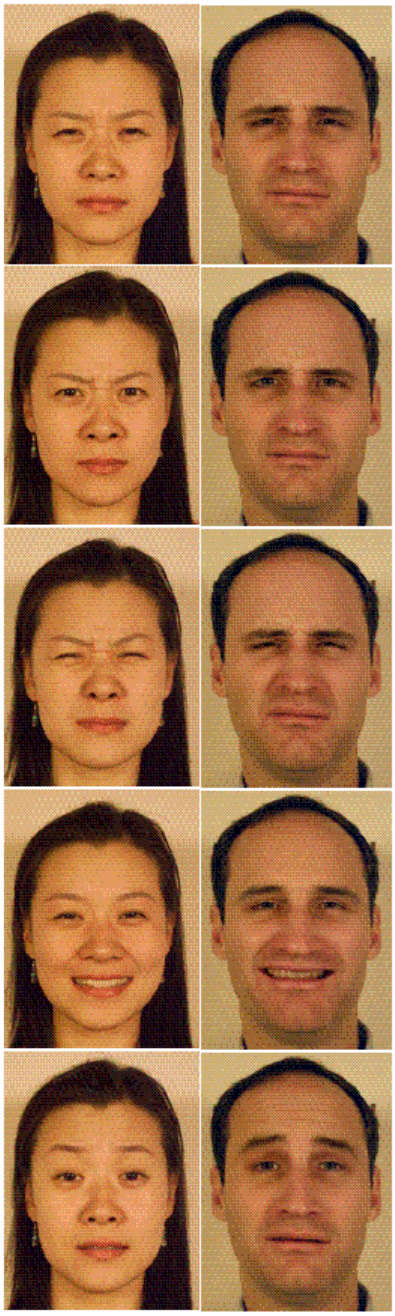


Figura 3-11: Resultados do mapeamento melhorado de expressões. As expressões da mulher são mapeadas no homem.

A Figura 3-12 ilustra algumas das expressões geradas pelo sistema de editor de expressões. É importante notar que cada uma dessas expressões possui uma geometria

diferente das imagens-exemplo. O sistema apresentado é capaz de produzir expressões faciais fotorealistas e convincentes.

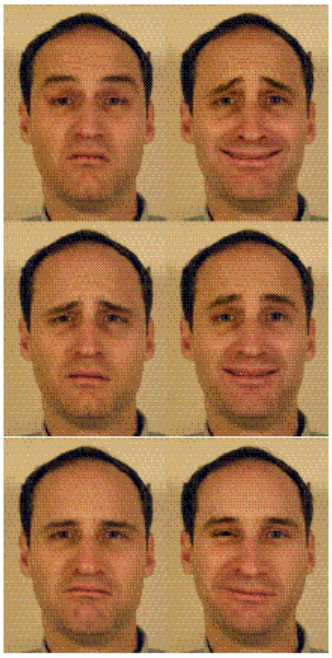


Figura 3-12: Expressões geradas pelo sistema editor de expressões.

Por fim, os autores capturaram um modelo de face 3D do homem através de um *laser scanner*. Foram também utilizados os movimentos dos pontos característicos para dirigir o movimento do vértice da malha 3D. Isso foi feito utilizando uma triangulação simples baseada em interpolação. Para cada quadro foi utilizada uma imagem de expressão sintetizada como textura para mapear na malha 3D. A Figura 3-13 ilustra os resultados.

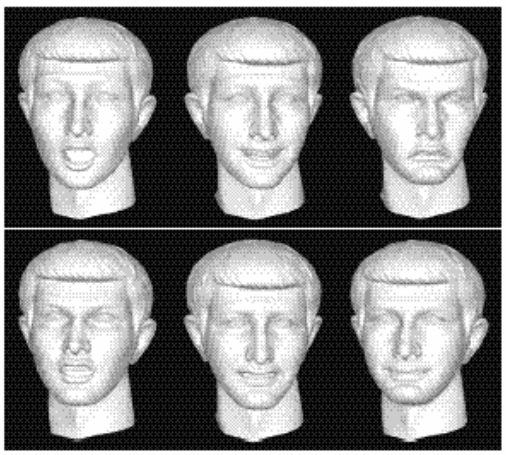


Figura 3-13: Resultados da síntese de expressão 3D.

3.14. CONCLUSÕES DO ARTIGO

O artigo apresentou um sistema de síntese de expressões faciais dirigido por geometria e uma técnica para inferência de pontos de controle. Os autores afirmam, em suas conclusões, que este é o primeiro sistema de mapeamento de expressões capaz de gerar detalhes das expressões, necessitando apenas dos movimentos dos pontos característicos do *performer*. O artigo também apresentou uma aplicação de editor de expressões onde o usuário pode manipular a posição geométrica dos pontos característicos e verificar, interativamente, as expressões faciais resultantes com aparência realista.

Como trabalhos futuros o artigo destaca:

- Melhorar a velocidade computacional através da aceleração do módulo de composição da imagem;
- Melhorar o alinhamento da imagem;
- Estender a técnica de síntese para permitir a síntese de expressões com várias poses dos exemplos;
- Conseguir gerenciar o movimento dos lábios durante a fala; e;
- Ser capaz de ter o mínimo de informações, tais como pontos característicos, poses e fonemas, do *performer* e automaticamente sintetizar animações faciais fotorealistas para o personagem alvo.

3.15. CONCLUSÕES PESSOAIS

i. Quem são os autores?

Qingshan Zhang: Pesquisador da Microsoft Research de Pequim, China (página pessoal não encontrada).

Zicheng Liu: Pesquisador da Microsoft Research de Redmond, WA. Fez doutorado em Princeton, terminando em 1996. Suas áreas de interesse são modelagem de face 3D, mapeamento de expressão facial, *face relighting*, comunicação e colaboração e processamento multi-sensor.

Baining Guo: Pesquisador Senior da Microsoft Research da Ásia, mestrado e doutorado em Cornell. Suas áreas de interesse são modelagem e *rendering* (incluindo síntese de textura, modelos de reflectância e sombreamento, real-time rendering, natural phenomena) e animação facial.

Harry Shum: Diretor de projetos da Microsoft Research na Ásia tem seu doutorado na área de Robótica na Universidade Carnegie Mellon. Áreas de interesse são visão computacional, computação gráfica, interação humano-computador, reconhecimento de padrão, aprendizado estatístico e robótica.

ii. O que o artigo resolve?

O artigo apresenta um sistema de síntese de expressões faciais baseado no mapeamento de expressões. A partir de imagens de entrada, o sistema é

capaz de montar a mesma expressão facial na face de um novo personagem sintetizado. O artigo também apresentou o desenvolvimento de um editor de expressões onde o usuário pode, interativamente, movimentar pontos característicos da face e ter, como resposta em tempo real, as novas expressões faciais que o próprio usuário está definindo. Por fim, o artigo apresentou uma técnica de inferência de movimento dos pontos característicos de um subconjunto baseado na abordagem de exemplos. Isto foi necessário porque as ferramentas de *track* existentes não conseguem extrair o número de pontos característicos necessários para manipulação no sistema.

iii. Qual a abordagem utilizada?

O artigo fez uso da técnica de animação baseada em performance, onde se têm pontos de controle da face que serão mapeados para sintetizar e montar as expressões faciais. Tendo consciência das limitações da técnica, o artigo propôs suprir essas limitações desenvolvendo uma nova técnica para inferência dos pontos de controle.

iv. Qual a classificação do artigo?

É um artigo de síntese de expressões faciais que busca resultados fotorealistas.

v. Quais foram as ferramentas utilizadas na implementação?

O artigo não aborda explicitamente as técnicas utilizadas na implementação do sistema de síntese de expressão facial. Uma das poucas coisas que o artigo menciona é o modelo de cabeça 3D que foi "escaneado" para se ter a malha 3D.

vi. Quais os problemas em aberto interessantes que o artigo apresenta ("ataca")?

Uma coisa interessante que o artigo aponta é o fato deles procurarem utilizar a animação dirigida por performance, aceitando a limitação de perder o realismo na expressão facial geral e ao mesmo tempo o artigo melhora a saída gerada trazendo realismo às expressões geradas.

Um outro aspecto interessante é o editor de imagens que, se funcionar como o artigo descreve, é uma ferramenta bastante interessante por ser interativa e com resposta em tempo real.

vii. O que não foi feito neste artigo que é interessante? (Quais os problemas em aberto interessantes que o artigo desperta e não ataca?)

Assim como no primeiro artigo apresentado [Bui2003], o trabalho preocupa-se com a produção de uma expressão facial realista. No entanto, para conseguir isto ele precisa abstrair da emoção facial casada com a fala. Os autores apontam a "emoção facial dirigida pela fala" como um dos possíveis trabalhos futuros.

Enfim, é possível afirmar que procurar trabalhar com animação facial de um personagem realista tendo as expressões faciais sincronizadas com a fala é um ponto bastante importante e que deve ser cada vez mais investigado.

4. Artigo III: "How Believable Are Real Faces? Towards a Perceptual Basis for Conversational Animation"

Este artigo [Cunningham2003] tem por objetivo determinar, de forma psicofisica, a ambigüidade e a credibilidade de expressões faciais. O artigo apresenta o resultado de um experimento desenvolvido com pessoas onde cada uma era solicitada para expressar uma determinada emoção (no caso foram consideradas as emoções *aggreement*, *disagreement*, *hapiness*, *sadness*, *thinking* e *confusion*). Outras pessoas, como avaliadores, foram solicitadas para tentar "adivinhar" as expressões.

Os resultados mostraram que as pessoas são capazes de identificar as expressões bem, apesar de existirem alguns padrões que levam a confusões. Os padrões específicos de confusões e taxas de corretude têm uma implicação forte na animação conversacional.

4.1. INTRODUÇÃO

Um grande número de movimentos da face e das mãos, com diferentes variedades, ocorre durante uma conversação. Alguns desses comportamentos não verbais são centrais tanto para o fluxo como para o significado da conversação [Cunningham2003].

Pesquisas mostram que pode ser excessivamente difícil produzir o padrão vocal necessário sem produzir o movimento facial acompanhado. Dessa forma, é possível afirmar que os movimentos faciais estão intimamente integrados com a mensagem falada. Isso leva a pensar que sinais visuais e auditivos devem ser tratados como unificados dentro da lingüística, e não como entidades separadas [Cunningham2003].

4.1.1. Trabalhos Relacionados

Foram poucos os trabalhos relacionados citados no artigo. Dentre os existentes, um trabalho interessante é o de Cassel et al. [Cassell2001]. O trabalho desenvolvido cria agentes que utilizam movimento de cabeça e dos olhos para ajudar o controle do fluxo de uma conversação (*help control turn-taking*).

4.1.2. Visão Geral do Experimento

Dentro de um contexto conversacional próprio, o artigo afirma que não somos capazes de determinar em 100% o que as expressões supostamente significam. Conhecer como os seres humanos percebem expressões conversacionais pode ser muito importante. Que movimentos faciais ou características distinguem uma expressão de outra? O que faz uma dada instância de uma expressão ser mais fácil de ser identificada do que uma outra instância da mesma expressão? O artigo [Cassell2001] buscou, através de experimentos, alcançar respostas e soluções para essas perguntas.

As "expressões universais", de acordo com Ekman [Ekman1971], são bastante conhecidas (Seção 1.3). Por outro lado, pouco se conhece a respeito das expressões não-afetivas que ocorrem durante uma conversação. Com o objetivo de maximizar o entendimento, a credibilidade e a eficiência dos agentes conversacionais, é de

fundamental importância conhecer os componentes necessários e suficientes para as várias expressões faciais.

Este artigo [Cunningham2003] iniciou sua pesquisa desenvolvendo um trabalho de base onde foram examinados em detalhes as seis expressões que o artigo assumiu como universais (*agreement*, *disagreement*, *happiness*, *sadness*, *thinking* e *confusion*), a fim de determinar quão diferentes elas são e as suas respectivas credibilidades. Utilizando um equipamento de gravação consistindo de 06 câmeras digitais sincronizadas (Seção 4.2), foram gravados 06 indivíduos executando essas expressões conversacionais (Seção 4.3). Os resultados do artigo (Seção 4.4) mostraram que as pessoas podem identificar essas expressões relativamente bem, embora haja alguns padrões sistemáticos de confusão.

4.2. EQUIPAMENTO DE GRAVAÇÃO

Para gravar as expressões faciais, um anel customizado de câmeras foi construído utilizando um sistema de gravação distribuído com seis unidades de gravação. Cada unidade de gravação foi composta por uma câmera digital de vídeo, um *frame grabber* e um computador, como ilustra a Figura 4-1.

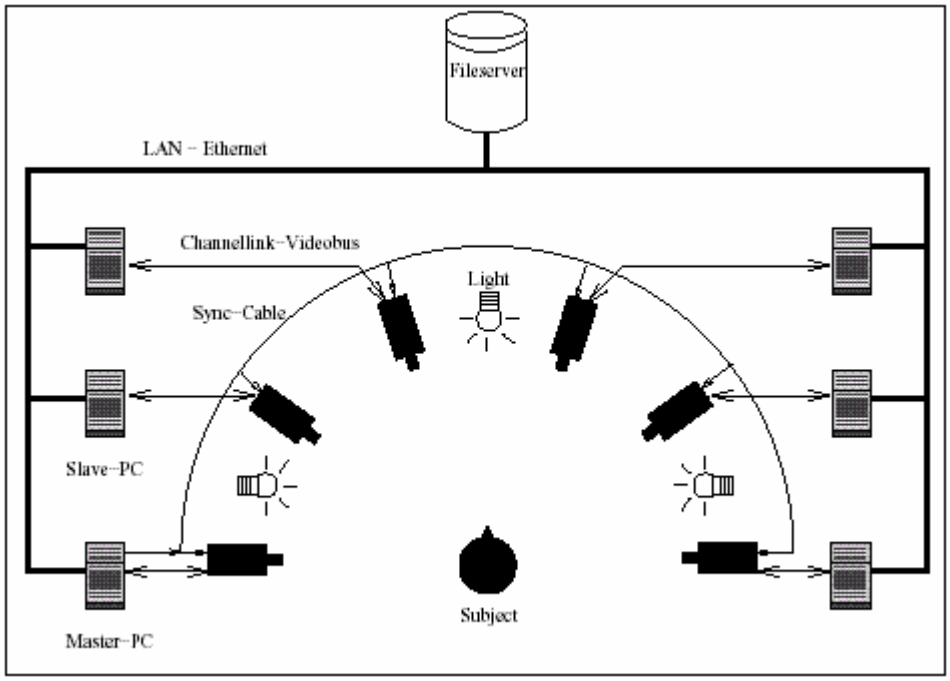


Figura 4-1: Esqueleto das 6 câmeras.

Cada unidade era capaz de gravar 60 *quadros/seg* de um vídeo não-comprimido, não-entrelaçado e completamente sincronizado em uma resolução PAL (768x576), armazenado em um formato CCD. As seis câmeras estavam dispostas em um semi-círculo em volta do objetivo da gravação em uma distância de aproximadamente 1,5 *m*. Os indivíduos foram filmados a 30 *quadros/s* e com um tempo de exposição de 3 *ms* com o objetivo de reduzir o borrar do movimento (*motion blur*) [Cunningham2003].

4.3. METODOLOGIA

Seis expressões foram gravadas de seis pessoas diferentes (três homens e três mulheres), levando a 36 seqüências de vídeo. As seis expressões foram de concordância, discordância, alegria, tristeza, pensamento e confusão, como ilustra a Figura 4-2.

Figura 4-2: As seis expressões: (a) *agreement* (concordância), (b) *disagreement* (discordância), (c) *hapiness* (felicidade), (d) *sadness* (tristeza), (e) *thinking* (pensativa) e (f) *confusion* (indecisa).

Para um processamento posterior nas gravações (por exemplo, reconstrução estérea), foram colocados marcadores pretos de *tracking* (rastreamento) na face em posições específicas. Depois da aplicação dos marcadores, cada ator (pessoa participante) foi centralizado em frente às seis câmeras. Cada pessoa era solicitada (induzida) a imaginar uma situação em que a expressão desejada ocorresse. O ator era primeiro filmado em sua expressão neutra, depois na expressão desejada e por fim na expressão neutra. Os atores não tinham restrições a respeito da dimensão ou quantidade de expressão que poderia ser feita, a única restrição era para que eles não falassem, a menos que fosse estritamente necessário. Esse procedimento foi repetido ao menos três vezes para cada emoção, e a melhor entre as repetições para cada pessoa era selecionada e editada.

Cada uma das 36 seqüências de vídeos foi mostrada para 10 pessoas diferentes (participantes) para o experimento psicofísico. O objetivo principal de um experimento psicofísico é o de sistematicamente examinar o relacionamento funcional entre as dimensões físicas (por exemplo, intensidade da luz) e as dimensões psicológicas (por exemplo, percepção de brilho) [Cunningham2003]. O trabalho examinou, de forma mais detalhada, o relacionamento funcional entre dimensões de alto nível (padrões de

movimentos faciais e percepção das expressões), e pode ser mais precisamente referenciado como um trabalho de médio a alto nível psicofísico.

As 36 expressões foram apresentadas de forma aleatória para cada um dos dez participantes. Os participantes foram submetidos a três tarefas. Na primeira tarefa, um participante via uma expressão, repetidamente, até responder. O intervalo entre uma repetição e outra era de 200 *ms* onde não era exibido nada. O participante precisava marcar (escolher) em um questionário de múltipla escolha o nome da expressão que ele tinha identificado. Além das seis expressões, havia uma sétima resposta com a opção "nenhuma das expressões acima", caso o usuário não conseguisse identificar a expressão. A segunda tarefa consistia em, uma vez identificada a expressão, indicar, em uma escala de um (pouco confiante) a cinco (completamente confiante), o quão confiante o participante estava sobre a sua escolha anterior. Por fim, a terceira tarefa consistia do participante responder de um (completamente falso) a cinco (extremamente verdadeiro) quão convincente a emoção era.

4.4. RESULTADOS E DISCUSSÃO

De uma forma geral, os participantes obtiveram sucesso em identificar as expressões. A Figura 4-3 ilustra uma tabela das respostas dos participantes. Para cada expressão, as respostas dos dez participantes foram unidas.

		Participants' Responses						
		Agreement	Disagreement	Happiness	Sadness	Thinking	Confusion	Other
Actual Expression	Agreement	**95%**	0%	2%	0%	0%	0%	3%
	Disagreement	0%	**85%**	2%	7%	0%	3%	3%
	Happiness	7%	2%	**73%**	3%	0%	5%	10%
	Sadness	0%	3%	0%	**82%**	5%	2%	8%
	Thinking	0%	3%	0%	2%	**73%**	20%	2%
	Confusion	0%	18%	0%	2%	5%	**73%**	2%

Figura 4-3: Matriz de confusão de identificação das respostas. O percentual de vezes que uma dada resposta foi escolhida (coluna) é mostrado para cada uma das expressões (linha).

A partir da tabela ilustrada em Figura 4-3 é possível afirmar que as expressões de *thinking* e *confusion* estão naturalmente interligadas (relacionadas). O artigo aponta duas justificativas para essa conclusão. A primeira é que quando uma pessoa está confusa em relação a alguma coisa, normalmente ela pára e pensa. O segundo ponto de justificativa para tal relacionamento deve-se ao fato que os atores não foram "treinados", e conseqüentemente eles podem não ter produzido a expressão corretamente [Cunningham2003].

A Figura 4-4 ilustra uma tabela em que claramente se vê que todas as expressões são potencialmente não-ambíguas mesmo sem estarem em um contexto. Como pode ser verificada, cada expressão foi reconhecida 100% ao menos para um ator, com exceção da expressão *thinking* que foi reconhecida no máximo em 90% na sua melhor vez.

Actual Expression	Actor 1	Actor 2	Actor 3	Actor 4	Actor 5	Actor 6
Agreement	100%	100%	100%	80%	100%	90%
Disagreement	90%	80%	100%	100%	90%	50%
Happiness	80%	50%	100%	60%	60%	90%
Sadness	40%	80%	90%	80%	100%	100%
Thinking	60%	90%	80%	70%	60%	80%
Confusion	90%	70%	50%	40%	90%	100%

Figura 4-4: Matriz de certeza do ator. A percentagem de vezes que uma dada expressão foi corretamente identificada é mostrada para cada um dos seis atores.

No geral, os participantes estavam completamente confiantes nas suas decisões, como ilustra a Figura 4-5.

Actual	Correct	False
Agreement	4.67	4.0
Disagreement	4.51	3.43
Happiness	4.44	4.23
Sadness	4.08	4.19
Thinking	4.30	3.96
Confusion	4.14	3.75

Figura 4-5: Taxa de confiabilidade: a média de confiança dos participantes em suas respostas é listada como uma função de se eles corretamente identificaram a expressão ou não. A confiabilidade foi classificada em uma escala variando de 5 pontos para "completamente confiante" a 1 ponto para "completamente não confiante".

Com relação ao resultado da terceira tarefa que eles desenvolveram, como ilustra a Figura 4-6, as expressões foram consideradas bastante convincentes, mas não completamente. Normalmente, os participantes achavam que uma expressão não era convincente justamente quando a identificavam incorretamente.

Actual	Correct	False
Agreement	3.81	3.25
Disagreement	3.96	2.79
Happiness	3.54	3.28
Sadness	3.26	3.69
Thinking	3.91	3.33
Confusion	3.67	3.73

Figura 4-6: Taxa de credibilidade (*believability*): a média de credibilidade é listada como uma função de se a expressão foi corretamente identificada ou não. A credibilidade foi julgada numa escala de 5 pontos para "completamente convincente" a 1 ponto para "completamente não convincente".

4.5. CONCLUSÕES DO ARTIGO

O artigo aponta como conclusões os seguintes tópicos:

- No geral, as seis expressões conversacionais são fáceis de serem identificadas, mesmo estando numa ausência completa de contexto conversacional;
- Existem alguns padrões de indecisão, mais notadamente nas expressões "*thinking"* e "*confusion"*, que foram algumas vezes interpretadas de forma errada; e;
- As pessoas geralmente estão confiantes que identificaram corretamente uma expressão mesmo quando de fato elas estavam erradas.

O artigo conclui que o fato de se ter tido alguma confusão nas respostas é conseqüência das expressões terem sido intencionalmente geradas. Existe uma evidência que deve ser levada em consideração em que durante uma conversação normal de humanos intencionalmente pode-se gerar várias expressões faciais, mas elas estão em sincronismo com uma porção auditiva da conversação [Cunningham2003].

Como conclusões finais, primeiramente, o artigo afirma que a criação de um modelo computadorizado 3D de uma cabeça que seja a duplicação física perfeita da cabeça real humana não significa que automaticamente as expressões geradas por essa cabeça sejam não-ambíguas e convincentes. Segundo, e talvez o mais importante, conclui-se que independente da razão de alguns indivíduos gerarem expressões faciais mais claras e convincentes, é mais interessante questionar que porções das seqüências de imagens de uma conversação levam as pessoas a ficarem confusas e que componentes aumentam o reconhecimento.

Por fim, como trabalhos futuros, o artigo aponta a possibilidade de desenvolver novos experimentos a partir das expressões presentes que foram gravadas de vários pontos de vista e com marcas de *tracking* na face. O grupo menciona a possibilidade de manipular os vídeos e assim fazer experimentos adicionais buscando elucidar que componentes são necessários e suficientes para se ter expressões conversacionais não-ambíguas e convincentes.

4.6. CONCLUSÕES PESSOAIS

i. Quem são os autores?

Douglas W. Cunningham: professor Dr. do *Max Plack Institute for Biological Cybernetics*. Ele coloca como sua principal área de interesse a integração da Computação Gráfica com a Psicofísica, afirmando que essas duas áreas, que são tradicionalmente separadas, possuem muito em comum e uma tem muito a doar para a outra. Outras áreas de interesse são: integração e adaptação *Visuomotor*, *Anorthoscopic Perception* (a.k.a., Aperture Perception), percepção binocular, Psicolingüística etc.

Martin Breidt: diretor técnico do *Max Plack Institute for Biological Cybernetics* responsável pela ponte entre a arte e a tecnologia. Coloca como suas áreas de interesse: computação gráfica, animação 3D, efeitos especiais, *scanning* 3D, captura de movimento, tecnologia de vídeo e *rendering* em tempo real.

Mario Kleiner: aluno de mestrado no *Max Plack Institute for Biological Cybernetics.*

Christian Wallraven: aluno de mestrado no *Max Plack Institute for Biological Cybernetics.*

Heinrich H. Bülthoff: Prof. Dr. Diretor do *Max Plant Institute for Biological Cybernetics Dept. Bülthoff***.** Coloca como objetivo de trabalho entender as funções do cérebro buscando analisar o comportamento humano para similar em ambientes virtuais.

ii. O que o artigo resolve?

O artigo apresenta um relato de experimentos feitos com seis indivíduos, fazendo, cada um, as seis expressões universais. Os vídeos foram analisados por 10 outros participantes que tentaram "adivinhar" qual era a expressão facial, quão conscientes eles estavam da sua resposta e qual convincente/realista era a expressão facial.

iii. Qual a abordagem utilizada?

Experimentos a partir da gravação de vídeos.

iv. Qual a classificação do artigo?

É um artigo de análise de expressões faciais.

v. Quais foram as ferramentas utilizadas na implementação?

Câmeras digitais para gravação de vídeo e marcadores para *tracking* nas faces dos atores.

vi. Quais os problemas em aberto interessantes que o artigo apresenta ("ataca")?

O artigo é bastante interessante, mas uma abordagem mais profunda seria ideal. Experimentos desse tipo são bastante interessantes de serem desenvolvidos, até como um mecanismo de validação do trabalho que está sendo desenvolvido.

vii. O que não foi feito neste artigo que é interessante? (Quais os problemas em aberto interessantes que o artigo desperta e não ataca?)

Faltou ao trabalho apontar quais são as principais partes de uma face que se têm maior e menor realismo em cada uma das seis expressões. Os autores colocam este ponto como um possível trabalho futuro.

5. Artigo IV: "Unsupervised Learning for Speech Motion Editing"

O artigo [Cao2003] discute o problema de edição de movimento facial gravado. Basicamente, o artigo apresenta uma técnica de aprendizado não-supervisionado baseada em análise de componentes independentes (ICA).

5.1. INTRODUÇÃO

A produção de uma animação facial de alta qualidade é um dos problemas mais desafiadores na animação por computador. Centenas de músculos contribuem para a geração de expressões faciais complexas e para a fala. A dinâmica desses músculos caracteriza-se por ser bastante difícil de se entender e, segundo relatos do artigo [Cao2003], até o momento, nenhum sistema desenvolvido é capaz de simular faces realistas em tempo real.

Existem diferentes abordagens que podem ser utilizadas no desenvolvimento de um sistema de animação facial realista. Este artigo faz uso da abordagem de gravar um conjunto representativo de movimentos e assim utilizar técnicas de aprendizado de máquina para estimar um modelo estatístico generativo (*generative statistical model*). O objetivo é então encontrar e "encaixar" um modelo que possibilite a re-síntese dos dados gravados.

Este artigo propõe uma técnica de aprendizado não-supervisionado baseado em análise de componentes independentes (ICA – *Independent Component Analysis*). Basicamente, a idéia é dividir (separar) os movimentos gravados em misturas lineares de fontes estatisticamente independentes. Essas fontes, chamadas de componentes independentes, fornecem uma representação dos dados com a semântica clara. A técnica é automática e não necessita de anotações dos dados.

5.2. TRABALHOS RELACIONADOS

Basicamente, o artigo classificou os seus trabalhos relacionados em dois grupos principais: síntese do movimento da face e análise do movimento.

5.2.1. Síntese do Movimento da Face

Para que a síntese da animação da fala seja feita, se faz necessário o conhecimento de uma biblioteca de formatos dos lábios (*visemas*) que possam ser mapeados em unidades de fala. Essa biblioteca pode ser construída de diversas formas. Uma opção é criar manualmente cada formato, mas para animação realista é preferível gravar esses formatos utilizando dados de um vídeo ou através da captura de movimento.

O trabalho *VideoRewrite* [Bregler1997] é um exemplo representativo do uso de uma base de dados de unidades audiovisuais.

Um outro trabalho interessante é o *Voice Pupperty* [Brand1999], que desenvolveu um mapeamento a partir da voz para a face através do aprendizado de um modelo da

dinâmica observada da face. O modelo leva em consideração a posição e a velocidade das características faciais e aprende a distribuição de probabilidade das diferentes configurações faciais.

Por fim, o trabalho *Multidimensional Morphable Model* [Ezzat2002] foi um modelo desenvolvido para a voz através do mapeamento da face focando no sincronismo dos lábios com a fala (*lip-sync*). O movimento da cabeça e da face superior foi tratado de forma *ad hoc*.

Enquanto que as técnicas citadas acima nos trabalhos relacionados ligados à síntese do movimento facial podem gerar um movimento de alta qualidade, elas normalmente não fornecem ao animador um controle intuitivo do estado emocional da *talking face*. No geral, essas técnicas focam no mapeamento do áudio com o sinal visual da fala e em efeitos que são gerados, como o de co-articulação[3]. Em contraste, o trabalho proposto neste artigo [Cao2003] apresenta o desenvolvimento de uma abordagem de aprendizado não-supervisionado que aprende dois mapeamentos diferentes: um entre o conteúdo fonético do sinal de áudio e o movimento da face, e outro entre o sinal de áudio e o conteúdo emocional da fala.

5.2.2. Análise do Movimento

Dentre os trabalhos de análise de movimento que o artigo apresenta, o trabalho [Chuang2002a] apresenta uma abordagem interessante para separar a fala visual em conteúdo e estilo (emoção). O método desenvolvido por eles faz uso de uma fatorização que produz um modelo bilinear que extrai emoção e conteúdo a partir de seqüências de vídeo dadas como entrada. Embora esse trabalho normalize os sinais, ele perde as informações temporais importantes e seus resultados são costurados por dados do vídeo.

A comunidade de reconhecimento de padrões tem desenvolvido uma quantidade significativa de trabalhos na análise de expressões faciais. As expressões são tipicamente baseadas no *tracking* (rastreamento) do movimento de elementos faciais particulares ou em características transientes como as rugas, como é possível verificar nos trabalhos [Black1995], [Essa1997] e [Lien1998]. Esses sistemas são bastante eficientes para o reconhecimento mas eles não se tornam muito claros quando se deseja utilizá-los para sintetizar ou editar o movimento facial [Cao2003].

O trabalho deste artigo emprega análise de componentes independentes (ICA – *Independent Component Analysis*) para extrair os modelos de conteúdo e de estilo de um conjunto grande de movimentos faciais gravados. O resultado consiste de componentes independentes que são a base para uma ferramenta de edição intuitiva de fala visual.

5.3. DECOMPOSIÇÃO FACIAL DO MOVIMENTO

Nessa seção o artigo apresenta uma visão geral da análise de componentes independentes (ICA). Ainda são abordadas a técnica de decomposição desenvolvida e a forma que são determinadas as semânticas dos componentes independentes resultantes.

[3] Co-articulação significa que o formato da boca usado para produzir um determinado fonema depende não apenas do fonema atual, mas também do fonema anterior e do fonema posterior.

5.3.1. Análise de Componentes Independentes

Análise de Componentes Infinitos (ICA) é uma técnica de aprendizado não-supervisionado [Hyvarinen2001]. Ela assume que um conjunto de variáveis aleatórias observadas possa ser expresso como combinações lineares de variáveis latentes independentes. De alguma forma, ela deconvolui (separa) os sinais gravados em um conjunto de variáveis aleatórias estatisticamente independentes.

Examinando a matemática de ICA, assumem-se n variáveis aleatórias $x_1,...,x_n$ observadas, cada uma sendo uma mistura linear de n variáveis latentes ou escondidas $u_1,...,u_n$, tal que $x_j = \sum_{i=1}^{n} a_{ji} u_i$, ou na forma matricial $x = Au$.

A equação $x_j = \sum_{i=1}^{n} a_{ji} u_i$ representa um modelo generativo: ela descreve como o dado gravado x é gerado pelas fontes u. As fontes u_i, que são chamadas de componentes independentes, não podem ser observadas diretamente. A matriz dos coeficientes A, chamada de matriz de misturas, é também desconhecida. ICA fornece um *framework* para estimar tanto A quanto u. Na prática, estimar A é suficiente, e uma vez a matriz conhecida, a sua inversa W pode ser aplicada para obter os componentes independentes: $u = Wx$.

Para estimar a matriz A a técnica ICA possui como vantagem o fato de seus componentes serem estatisticamente independentes. A chave de estimar o modelo ICA é a não-gaussianidade. De acordo com o Teorema do Limite Central, a soma de duas variáveis aleatórias independentes usualmente possui uma distribuição próxima a uma distribuição gaussiana. A idéia é então interativamente extrair variáveis aleatórias a partir dos dados gravados de forma que sejam "não-gaussianos quanto possível". O artigo afirma que como a não-gaussianidade é medida está fora de seu escopo. Nos experimentos do artigo foi utilizada uma implementação pública disponível, chamada de *FastICA*[4].

5.3.2. Pré-Processamento

Antes da aplicação da técnica ICA, os dados precisam passar por uma fase de pré-processamento que consiste de dois passos: centralização (*centering*) e *whitening*.

A centralização desloca os dados em torno de sua média de tal forma que as variáveis aleatórias resultantes tenham média zero. O *whitening* transforma o conjunto centralizado de variáveis observadas em um conjunto de variáveis não-correlacionadas. Análise de Componentes Principais (PCA – *Principal Component Analysis*) pode ser usada para executar essa transformação. Depois da fase de pré-processamento, a equação $x = Au$ assume a forma $x = E\{x\} + PAu$, onde $E\{x\}$ é a esperança de x e P é uma matriz *nxm* obtida da aplicação de PCA ao dado centralizado; *m* é o número de componentes principais.

[4] FastICA em http://www.cis.hut.fi/projects/ica/fastica/.

5.3.3. PCA *versus* ICA

PCA e ICA são técnicas estatisticamente relacionadas. Ambas fornecem uma decomposição linear dos dados amostrados. A diferença fundamental é que PCA assume que as variáveis latentes são não-correlacionadas e ICA assume que elas são independentes (variáveis aleatórias independentes são sempre não-correlacionadas, mas o inverso não é verdade).

O objetivo do PCA é encontrar uma seqüência de variáveis (componentes) aleatórias não-correlacionadas onde cada variável cobre o máximo possível da variância dos dados. A seqüência resultante é ordenada pela cobertura decrescente da variância. Por essa razão, PCA é geralmente uma técnica eficiente de compressão: guardando os primeiros poucos componentes a maioria da variância do dado pode ser coberta.

Os componentes independentes produzidos pela ICA fornecem um mecanismo de separação entre as fontes que são assumidas serem independentes, ao invés de um mecanismo de compressão.

5.3.4. Aplicação do Movimento Facial

Aplicar ICA ao movimento facial gravado é uma tarefa direta. O movimento é representado como um conjunto de séries do tempo $x_i(t)$ que captura as coordenadas euclidianas das marcas dos movimentos no tempo. Cada uma dessas séries pode ser pensada como amostras de variáveis aleatórias x_i, usando $x = E\{x\} + PAu$.

Essa decomposição resulta em um conjunto de componentes independentes que possui uma interpretação intuitiva.

5.4. INTERPRETAÇÃO DE COMPONENTES INDEPENDENTES

O trabalho apresentado no artigo [Cao2003] decompõe movimentos relacionados à fala em um conjunto de fontes (origens) que podem ser claramente interpretadas e manipuladas para o propósito de edição. Em particular, o trabalho separou os dados em componentes de estilo e componentes de conteúdo. Componentes de estilo estão diretamente relacionadas com expressividade ou emoção, enquanto componentes de conteúdo estão relacionadas com a parte do movimento responsável pela formação da fala.

5.4.1. Número de Componentes Independentes

Antes de aplicar ICA sobre os dados, foi necessário determinar o número de componentes a serem extraídos. Não existe nenhuma regra clara para essa quantidade.

Na prática, como relatado em [Cao2003], o passo de pré-processamento *whitening* reduz a dimensão do dado e determina o número de componentes independentes. Na maioria dos experimentos, guardar os componentes que cobriam cerca de 95% a 98% da variância fornecida era suficiente.

5.4.2. Emoção

A emoção da face de um ator foi gravada enquanto ele estava pronunciando um conjunto de sentenças por várias vezes, cada vez expressando uma emoção diferente. Supõe-se (x^i, y^i) como sendo p pares de movimentos que correspondem à mesma sentença com duas expressões diferentes. Aplicando ICA para cada par de movimento na base de dados, irá resultar em pares que correspondem a conjuntos de componentes independentes (u^i, v^i).

É esperado que componentes independentes relacionados à emoção sejam significantemente diferentes entre os dois movimentos de fala com o mesmo conteúdo mas diferentes emoções. Com o objetivo de validar essa propriedade, cada par de movimentos correspondentes foi alinhado usando *Dynamic Time-Warpingi* (DTW) [Cao2003]. Sejam (u'^i, v'^i) os componentes independentes de dois movimentos alinhados depois do *warping*. A diferença entre eles é computada através de o erro de média da raiz quadrada (RMS - *Root Mean Square*) dado por

$$d_{emotion,j} = \left(\frac{1}{\sum q_i} \left(\sum_{i=1}^{p} \left(\sum_{k=1}^{q_i} \left(u'^i_j(t_k) - v'^i_j(t_k) \right)^2 \right) \right)^{\frac{1}{2}} \right)$$

onde q_i é o número de amostras alinhadas no tempo para o par i. A distância $d_{emotion,j}$ é calculada de tal forma que ela deve ser grande se o componente j estiver relacionado à emoção.

A Figura 5-1 ilustra o gráfico dos valores de $d_{emotion,j}$ para 6 componentes independentes estimados a partir de 32 pares de sentenças das emoções "frustado e feliz" (*Frustated* e *Happyi*). Existe um pico na terceira componente que indica, justamente, que essa é uma componente relacionada com variações emocionais. Os outros componentes participam em menor grau no componente emocional dos movimentos. Isso mostra que o movimento de fala não pode ser estritamente separado em componentes estatisticamente independentes. A abordagem do artigo possui uma boa aproximação.

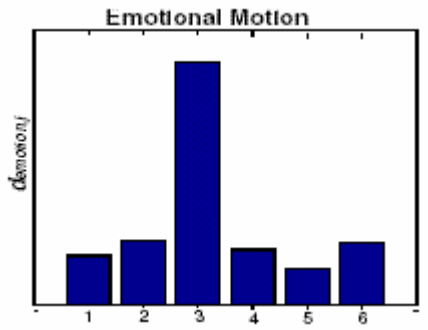


Figura 5-1: Classificação dos componentes independentes para emoção. O eixo horizontal representa o índice dos componentes independentes e o eixo vertical indica a métrica da distância $d_{emotion,j}$.

5.4.3. Conteúdo

O artigo definiu conteúdo como sendo parte do movimento associado com a formação da fala independentemente da expressão. Para esse caso foram considerados apenas os movimentos das marcações na área da boca (12 marcas).

Definindo a métrica da distância entre dois movimentos que foram reconstruídos usando dois subconjuntos de componentes independentes *A* e *B*, tem-se:

$$d_{mouth}(x_A, x_B) = \left(\frac{1}{q} \sum_{k=1}^{q} \left(\frac{1}{r} \sum_{l=1}^{r} \left(x_A^l(t_k) - x_B^l(t_k) \right)^2 \right) \right)^{\frac{1}{2}}$$

onde x_A e c são as emoções reconstruídas usando o componente do subconjunto *A* e *B*, respectivamente, e q é o número de marcas consideradas para a região da boca (12 marcas).

Reconstruindo o movimento das marcas da boca utilizando todos os componentes independentes tem-se x_{all}. Com o objetivo de avaliar quanto cada componente independente i contribui para o movimento da boca, a seguinte métrica é computada: $d_{mouth,i} = d_{mouth}(x_{E\cup\{i\}}, x_{all}) - d_{mouth}(x_E, x_{all})$, onde E é o subconjunto de componentes independentes responsáveis pela emoção e x_E é o movimento da marca reconstruída a partir do subconjunto *E*.

A equação $d_{mouth}(x_A, x_B) = \left(\frac{1}{q} \sum_{k=1}^{q} \left(\frac{1}{r} \sum_{l=1}^{r} \left(x_A^l(t_k) - x_B^l(t_k) \right)^2 \right) \right)^{\frac{1}{2}}$, $d_{mouth,i}$ quantifica a influência do componente independente i no movimento da boca. O quão grande esse valor absoluto seja, maior será a influência do componente i sobre o movimento da boca. A Figura 5-2 mostra o valor de $d_{mouth,i}$ para os seis componentes independentes. É possível verificar quão grandes são $d_{mouth,1}$, $d_{mouth,4}$ e $d_{mouth,5}$ comparados com os restantes dos componentes. Visualmente é possível verificar que o movimento $x_{\{1\}\cup\{4\}\cup\{5\}}$ reconstruído usando os componentes 1, 4 e 5 captura grande parte do movimento da fala.

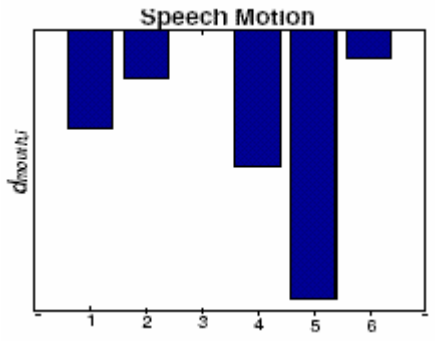


Figura 5-2: Classificação dos componentes independentes para a fala. O eixo horizontal representa o índice dos componentes independentes e o eixo vertical indica a métrica da distância $d_{mouth,i}$.

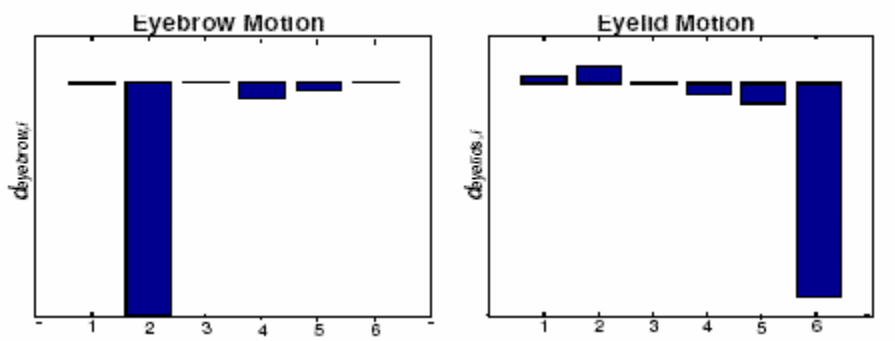
5.4.4. Movimento de Piscar das Pálpebras e Movimentos da Sobrancelha Não-Emocionais

Experimentos mostraram que alguns componentes independentes não podem ser associados nem com a emoção nem com o conteúdo. Esses componentes podem ser classificados em dois outros grupos: movimento de piscar e movimento não-emocional da sobrancelha (movimentos de estresse e ênfase da fala).

Com o objetivo de identificar os componentes relacionados a esses dois tipos de movimentos foi utilizado o mesmo método aplicado para encontrar os componentes relacionados com o conteúdo. Foram definidos $d_{eyebrown}$ e $d_{eyelids}$ levando em consideração marcas nas sobrancelhas e nas pálpebras, respectivamente.

A Figura 5-3 (a) ilustra o valor da métrica da distância $d_{eyebrown,i}$ para os 6 componentes independentes. É importante observar quão maior é a distância $d_{eyebrown,2}$ comparado com a distância dos outros componentes. Claramente o componente 2 captura a maioria dos movimentos da sobrancelha.

De forma similar, a Figura 5-3 (b) ilustra a métrica da distância $d_{eyelids,i}$ para os 6 componentes independentes. Nesse caso $d_{eyelids,6}$ domina, comparado com o restante dos componentes, indicando assim que o componente 6 captura grande parte do movimento das pálpebras.



(a) (b)

Figura 5-3: Classificação dos componentes independentes para a sobrancelha em (a) e para a pálpebra em (b). O eixo horizontal representa o índice dos componentes independentes e o eixo vertical indica a métrica da distância $d_{eyebrown,i}$ em (a) e $d_{eyelids,i}$ em (b).

5.5. EDIÇÃO

Baseado na decomposição proposta, o artigo apresenta o desenvolvimento de uma ferramenta de edição de movimento facial. Essa ferramenta permite que o usuário interativamente modifique a aparência do conteúdo emocional da fala visual.

As operações permitidas na ferramenta de edição foram: translação, cópia e substituição, e cópia e adição. Após a alteração de parâmetros do modelo devido à edição

do movimento, era necessário utilizar a representação ICA para re-sintetizar o movimento.

O modelo ICA pode ser escrito como $x = E\{x\} + PAu$, onde existem três parâmetros que podem ser modificados: a média $E\{x\}$, a matriz de mistura PA e os componentes independentes u.

A operação de translação permite que uma emoção seja modificada diretamente. A idéia consiste em estimar os valores extremos e transladar a série do tempo responsável pela emoção entre esses dois extremos. Com essa técnica é possível modificar a emoção continuamente entre duas emoções presentes no conjunto de treinamento. A edição pode ser expressa através da equação $x = E\{x\} + PA(u + \alpha e_E)$, onde α é um escalar que determina a quantidade de translação no componente emocional e e_E é um vetor na base canônica da matriz de mistura do ICA, que corresponde ao componente emocional.

A operação de copiar e substituir tem o propósito de substituir o componente emocional de um movimento por um componente emocional de um movimento diferente sem trocar o conteúdo (fala do movimento) do movimento original. Para fazer essa substituição, é necessário substituir a série do tempo que corresponde ao componente emocional u_1 no espaço ICA pelo componente emocional do segundo movimento u_2. Essa manipulação pode ser representada pela equação $x = E\{x\} + PA\left(u_1 + \left((u_2 - u_1)^T e_E\right)e_E\right)$.

A operação de copiar e adicionar tem o objetivo de permitir a adição de um componente emocional que não está presente no movimento original. Considerando u_1 e u_2 os componentes emocionais dos dois movimentos, para adicionar o componente emocional do movimento 1 ao movimento 2 a seguinte operação é executada: $x = E\{x\} + (PA)_1 u_1 + (PA)_2\left(\left(u_2^T e_E^2\right)e_E^2\right)$, onde $(PA)_1$ e $(PA)_2$ são as matrizes de mistura dos dois movimentos e e_E^2 é o vetor na base canônica da matriz do componente emocional.

5.6. Resultados, Conclusões do Artigo e Trabalhos Futuros

O artigo aponta como resultado que o ICA pode ser utilizado para decompor o movimento facial relativo à fala em componentes significativos. Ele indica a URL http://www.cs.ucla.edu/~abingcao para constatação e verificação dos resultados.

Como principais contribuições o artigo aponta a proposta de utilizar uma técnica de aprendizado não-supervisionado baseado em ICA para extrair parâmetros significativos de um conjunto de movimentos faciais gravados. Uma segunda contribuição que o artigo aponta, e também considera bastante significativa, foi o fato de mostrar que movimentos faciais podem levar facilmente a decomposições lineares, abstraindo-se da complexidade associada com o sistema de controle (o cérebro) e os mecanismos responsáveis por esses movimentos.

Como trabalhos futuros, o artigo aponta o desejo de analisar movimentos faciais mais complexos, por exemplo, movimentos faciais com emoções sem restrições. Seria bastante interessante verificar se o ICA conseguiria extrair os componentes emocionais desses dados. Um outro trabalho futuro interessante que o artigo aponta é a extensão do

conjunto de operações de edição nos componentes independentes, fazendo uso de técnicas de processamento de sinal.

5.7. CONCLUSÕES PESSOAIS

i. Quem são os autores?

Yong Cao: Aluno de doutorado da UCLA. Tem como orientador Petros Faloustsos e como co-orientador do Frédéric Pighin. Áreas de pesquisa são computação gráfica, animação facial e animação do corpo, síntese de movimento baseada em busca e aprendizado de máquina para síntese de movimento.

Petros Faloutsos: Professor Assistente do Departamento de Ciência da Computação na universidade UCLA. Áreas de interesse: computação gráfica, animação baseada em movimentos físicos, robótica e biomecânica.

Frédéric Pighin: Pesquisador senior do Institute for Creative Technologies (University of Southern California) onde ele é o gerente do Computer Animation Lab. Tem como objetivo desenvolver técnicas para sintetizar atores virtuais foto-realistas, tendo como interesse tanto a representação de atores sintéticos para rendering e animação, quanto o desenvolvimento de ferramentas para permitir que artistas digitais criem seus personagens humanos com aparência convincente.

ii. O que o artigo resolve?

O artigo apresenta uma ferramenta de edição da fala estando ela relacionada a movimentos faciais. O sinal de fala é separado em duas "componentes" principais: o conteúdo da fala propriamente dita e a emoção associada com a fala em questão. A ferramenta de edição faz uso de uma técnica de aprendizado não-supervisionado baseada na análise de componentes independentes (ICA). Através da ICA é possível separar o conteúdo e a emoção e assim conseguir verificar de qual "tipo" aquele componente é para assim fazer as operações de edição de forma correta.

iii. Qual a abordagem utilizada?

O artigo faz uso de uma técnica de estatística que á a análise de componentes independentes para extrair as informações desejadas.

iv. Qual a classificação do artigo?

É um artigo de análise tanto de expressões faciais quanto da fala, onde ele consegue separar esses dois componentes de forma eficiente para modificá-los independentemente, mas mantendo a relação existente na fala e na emoção em questão.

v. Quais foram as ferramentas utilizadas na implementação?

O artigo não comenta, mas um ferramental estatístico e matemático foi bastante utilizado.

vi. Quais os problemas em aberto interessantes que o artigo apresenta ("ataca")?

Achei o artigo bastante interessante. Acho que as técnicas de estatística que vêm sendo utilizadas nessa área de animação facial têm enriquecido bastante os resultados dos trabalhos. Foi uma técnica nova para tentar solucionar um problema ainda em aberto, que é justamente essa separação da emoção e do conteúdo, principalmente na região da boca.A ferramenta de edição desenvolvida também foi algo bem interessante, principalmente no que se refere a permitir uma interatividade do usuário para transladar uma expressão, substituir ou adicionar uma nova expressão. Um outro comentário importante do artigo é que ele propõe fazer como trabalho futuro o uso de expressões "compostas", ou seja, o personagem não estará "feliz", a sua emoção de felicidade vai ter uma característica quantitativa, ele poderá estar "muito feliz", "mais ou menos feliz" ou "pouco feliz", como também será possível compor emoções. É um trabalho a ser explorado (não trabalhar apenas com emoções puras).

vii. O que não foi feito neste artigo que é interessante? (Quais os problemas em aberto interessantes que o artigo desperta e não ataca?)

Apesar da técnica de análise de componentes independentes ter sido explorada de maneira interessante, fiquei com vontade de ver o seu detalhamento na área da região da boca. O artigo apenas coloca na boca a contribuição do conteúdo, faltando adicionar a contribuição da emoção. Esse casamento é um tema que vem sendo pouco explorado e seria interessante verificar como uma abordagem estatística poderia ajudar nisso.

6. Artigo V: "Learning Controls for Blend Shape Based Realistic Facial Animation"

O artigo [Joshi2003] propõe uma técnica automática de segmentação fisicamente motivada para aprender os controles e os parâmetros de uma animação diretamente do conjunto de *blend shapes*. Essa técnica será utilizada, de forma eficiente, tanto para animação baseada na captura de movimento quanto para animação *keyframing*.

6.1. INTRODUÇÃO

A animação facial precisa de um modelo deformável da face para poder expressar a grande quantidade de configurações faciais relacionadas à fala e às emoções. Basicamente, existem duas formas tradicionais de criar modelos faciais deformáveis [Joshi2003]: usando um modelo físico (*physically-based model*) ou um modelo baseado em *blend shape*[5]. A primeira abordagem foi apresentada na Seção 1.2.5 e por não ser o modelo abordado neste artigo não será reapresentada.

O modelo *blend shape*, ao invés de ficar precisamente preocupado com a mecânica da face (como acontece no modelo físico), considera cada expressão facial como uma combinação linear de algumas expressões faciais selecionadas que são os *blend shapes*. Através da variação dos pesos da combinação linear, um conjunto completo de expressões faciais pode ser expresso necessitando de pouca computação.

Hoje em dia existem várias opções de se criar *blend shapes*. Um artista digital habilidoso pode deformar a malha em diferentes formatos canônicos necessários para cobrir o número de expressões. Alternativamente, os *blend shapes* podem ser diretamente escaneados de um ator real ou de um modelo de argila.

O restante desta seção destina-se a apresentar alguns dos principais trabalhos relacionados mencionados no artigo [Joshi2003] e uma visão geral das principais contribuições do mesmo.

6.1.1. Trabalhos Relacionados

Interpolação *blend shape* pode ser apresentada como um dos primeiros trabalhos desenvolvidos por Parke [Parke1974]. A idéia original desenvolvida por ele foi rapidamente estendida para segmentação da face onde as regiões eram combinadas individualmente. Tradicionalmente, essas regiões eram definidas manualmente.

Uma forma simples era segmentar a face em duas regiões: a região superior que era destinada para as características emocionais da face, e a região inferior que representava os movimentos da fala. Embora essa separação fosse, e ainda seja, normalmente usada, ela não está próxima de uma solução "ideal" já que ela não reflete as interdependências que há entre essas duas regiões.

Tem existido pouca pesquisa na área de manipulação interativa de modelos *blend shape* com exceção do trabalho desenvolvido por Pighin et al [Pighin1998]. O trabalho

[5] Ao longo do texto será utilizado o termo em inglês *blend shape*, por não saber da tradução técnica e por não achar "formato combinado" ou "mistura de formas" uma tradução plausível.

deles descreve um sistema de animação *keyframe* que faz uso de uma palheta de expressões faciais ao longo de uma interface de pintura a fim de atribuir pesos às combinações. O sistema fornece ao animador a liberdade de atribuir os pesos de combinação na granularidade de um vértice. O artigo afirma que essa liberdade pode se tornar um inconveniente por não levar em consideração as limitações físicas da face e assim tornar mais difícil a criação de faces realistas.

O sistema proposto neste artigo [Joshi2003] afirma ser um pouco diferente: ele não respeita a mecânica da face através da análise das propriedades físicas dos dados. Comparativamente, o sistema deste artigo afirma ser mais intuitivo e capaz de gerar expressões faciais plausíveis.

Choe et al [Choe2001] desenvolveu um trabalho interessante de mapeamento de movimento em um conjunto de *blend shapes*. O ponto fraco do trabalho deles foi o desenvolvimento de uma segmentação manual da face, enquanto o sistema deste artigo aprende essa segmentação a partir dos dados (de forma automática). Basicamente, este trabalho tem o objetivo de segmentar uma malha 3D que seja uma combinação linear de malhas amostradas.

6.1.2. Contribuição e Visão Geral

O artigo [Joshi2003] tem como foco os problemas de parametrização e controle de modelos *blend shapes*. O artigo apresenta o desenvolvimento de uma técnica automática que extrai um conjunto de parâmetros a partir de um modelo *blend shape*. Ao invés de derivar o mecanismo de controle da biomecânica da face, o artigo aprende esse controle diretamente dos dados disponíveis. Essa solução é específica para os *blend shapes* processados e reflete as idiossincrasias presentes nos dados.

O artigo também demonstra a utilidade desses parâmetros em duas técnicas de animação: *motion capture* (captura de movimento) e *keyframing*. Por fim, o artigo apresenta um novo algoritmo de *rendering* para modelos *blend shapes*.

6.2. MODELO DA FACE BLEND SHAPE

O artigo define um modelo de face *blend shape* como sendo uma combinação linear convexa de *n* vetores de base, cada vetor sendo um dos *blend shape*. Cada *blend shape* é um modelo de face que inclui geometria e textura, e todas as malhas *blend shapes* para um dado modelo possuem a mesma topologia. As coordenadas de um vértice V pertencente ao modelo *blend shape* podem ser escritas através da equação $V = \sum_{i=1}^{n} \alpha_i V_i$ onde os escalares α_i são os pesos combinados, V_i é a posição do vértice no *blend shape* i e n é o número de *blend shapes*. Os pesos devem satisfazer a restrição de convexividade onde $\alpha_i \geq 0, \forall i$ e $\sum_{i=1}^{n} \alpha_i = 1$ que indica uma soma igual a um para a invariância rotacional e translacional.

De forma similar, a textura em um ponto particular de um modelo *blend shape* é uma combinação linear (*alpha blending*) das texturas do *blend shape* com o mesmo peso de combinação usado para a geometria.

6.2.1. Modelo Físico

Um dos modelos físicos mais simples para deformação de objetos é o de elasticidade linear. A deformação de um objeto é medida pelo campo de deslocamento *d* entre a posição corrente de cada ponto e sua posição final. A equação que governa o movimento de um modelo elástico linear é a formulação de Lamé dada por $\rho a = \lambda \Delta d + (\lambda + \mu)\nabla(\nabla \cdot d)$.

No contexto deste artigo, *d* é o deslocamento do vértice a partir de sua posição na face neutra, ρ é a média da densidade de massa da face, *a* é a aceleração do vértice e λ e μ são coeficientes de Lamé que determinam o comportamento do material [Joshi2003].

6.2.2. Segmentação

O valor discreto do laplaciano na equação $\rho a = \lambda \Delta d + (\lambda + \mu)\nabla(\nabla \cdot d)$ é computado para cada vértice de cada *blend shape* não-neutro (expressão não-neutra) e a magnitude do vetor resultante é calculada. Esse cálculo fornece um mapa de deformação para cada expressão. O mapa resultante, como ilustra Figura 6-1, mede a quantidade máxima de deformação local que o modelo de face do artigo possui para os *blend shapes* utilizados. Uma segmentação mais rápida pode ser obtida através da divisão desse mapa em regiões com menor deformação e em regiões com maior deformação. O limiar (*threshold*) para essa divisão pode ser escolhido através da equação $threshold = D[nt]$, onde *D* é o vetor ordenados dos valores de deformação, *n* é o tamanho do vetor e *t* é um escalar variando de 0 a 1.

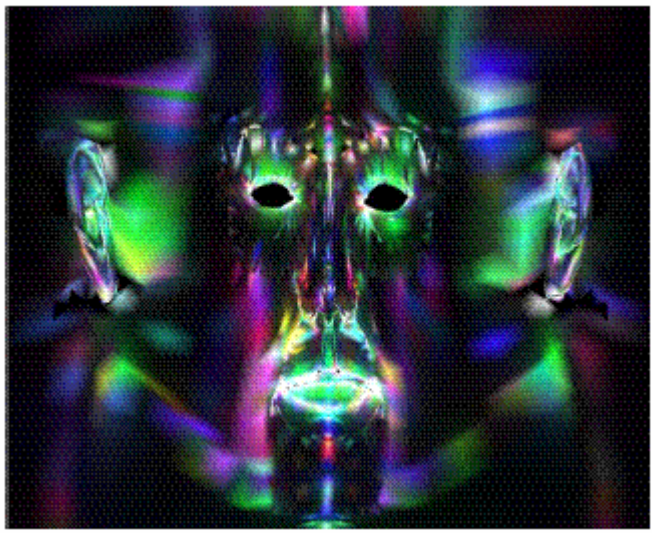
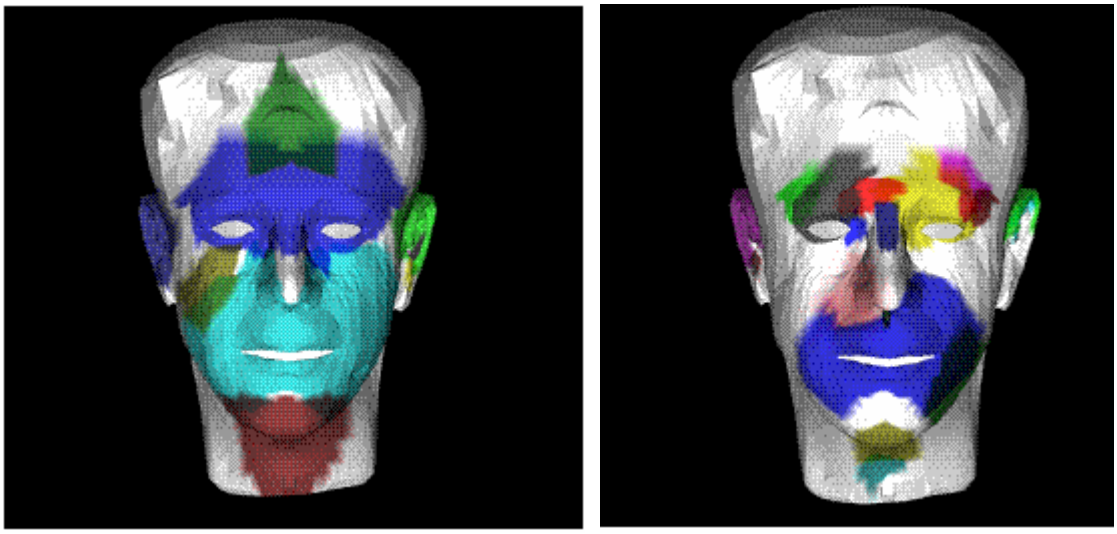


Figura 6-1: Mapa de deformação das regiões geradas automaticamente (a deformação nas direções X, Y e Z é expressada como uma tripla RGB, respectivamente).

Em síntese, a etapa de segmentação é compreendida pelos seguintes passos: primeiro todos os valores da deformação são ordenados e depois se obtém a deformação na posição que é uma função do número de valores. Por exemplo, para utilizar as regiões na Figura 6-2 (a) e Figura 6-2 (b), foram utilizados, respectivamente, um limiar de *t=0.25* e *t=0.75*.



(a) (b)

Figura 6-2: Regiões automaticamente geradas: (a) segmentação utilizando um limiar baixo (*threshold=0.25*) e (b) segmentação utilizando um limiar alto (*threshold=0.75*).

6.3. ANIMAÇÃO COM CAPTURA DE MOVIMENTO

O artigo [Joshi2003] expressa o movimento na captura de movimento de um dado utilizando o modelo *blend shape*. Assume-se que o movimento (ou a posição por quadro) de uma marca de movimento pode ser expressa como uma combinação linear de pontos correspondentes nos *blend shapes*. Nomeando, tem-se $M_j = \sum_{i=1}^{n} \alpha_i V_{ij}$, onde M_j é uma posição da face cujo o movimento foi gravado, V_{ij} é a posição correspondente no *blend shape* i, m é o número de marcas de movimento e n é o número de *blend shapes*.

Dada tais equações, foi possível encontrar os pesos α_i. A equação acima foi remodelada como um problema de minimização, onde foi preciso minimizar a soma da diferenças $\sum_{j=1}^{m}\left[M_j - \left(\sum_{i=1}^{n}\alpha_i V_{ij}\right)\right]^2$. Todo o sistema é um sistema de equações lineares onde os α_i desconhecidos são os pesos na combinação *blend shape*. Utilizando uma solução de programação quadrática interativa foi possível obter os valores ótimos dos pesos α_i combinados.

Para produzir uma malha animada que segue o movimento mais precisamente, o artigo completou a projeção da base do *blend shape* através de uma translação dos vértices na malha por um resíduo $\left(M_j - \sum_{i=1}^{n}\alpha_i \cdot V_{ij}\right)$. As coordenadas finais V_j de um

vértice em uma face são construídas usando $V_j = P_j + RBF(P_j)$, onde P_j é a projeção no conjunto de *blend shape* $P_j = \sum_{i=1}^{n} \alpha_i V_{ij}$ e $RBF(P_j)$ é o resto interpolado em cada vértice P_j, $RBF(P_j) = \sum_{i=1}^{m} \exp(-\|M_i - P_j\|) C_i$.

Ao invés de resolver o sistema de equação acima para o modelo inteiro, o artigo solucionou para cada região criada a partir do processo de segmentação automática. Com isso, foi possível ter um controle localizado sobre toda a malha da face e, conseqüentemente, ter resultados melhores nas restrições espaciais. Segundo o artigo, a solução local também possibilitou expressar um grande número de movimentos utilizando apenas um número limitado de *blend shapes* (no caso deles 10).

O problema de minimização citado foi construído para cada quadro e para cada região e, assim, obtido os pesos da combinação. Os mesmos pesos foram utilizados para todos os vértices de uma região para obter as novas posições que casavam com o movimento. Assim, para cada quadro do movimento, foi possível solucionar o problema de minimização para obter os pesos da combinação e conseqüentemente a malha da face seguiu o movimento capturado do dado.

6.4. EDIÇÃO KEYFRAME

Através do modelo *blend shape* foi possível interativamente construir malhas de faces que podem ser usadas como *keyframes* (quadros-chave) em uma ferramenta de animação facial baseada em *keyframing*.

Criar um quadro-chave (*keyframe*) é similar a produzir um quadro em uma seqüência de captura de movimento onde se faz necessário especificar os pontos de controle (marcas), os seus respectivos mapeamentos e restrições espaciais (por exemplo, as posições das marcas). Na interface desenvolvida neste artigo, o usuário podia interativamente especificar todas as características descritas acima através do clicar e do arrastar com o *mouse* no modelo facial. Assim como no processo de animação por captura de movimento, através da equação $\sum_{j=1}^{m} \left[M_j - \left(\sum_{i=1}^{n} \alpha_i V_{ij} \right) \right]^2$, foi construído um problema de minimização usando as restrições interativamente especificadas e foram obtidos os pesos combinados.

Foi possível segmentar o modelo *blend shape* em regiões usando a técnica de segmentação automática descrita anteriormente. Com o objetivo de permitir uma edição *keyframe* em vários níveis de detalhes, foi construída uma hierarquia de regiões. Primeiro essa hierarquia foi criada utilizando o algoritmo de segmentação com um valor alto de limiar de forma a gerar regiões pequenas e localizadas.

A Figura 6-3 ilustra uma seqüência de manipulações executadas sobre um quadro. A edição sucessiva do quadro foi feita aumentando o nível de detalhe para refinar a expressão facial de maneira localizada.

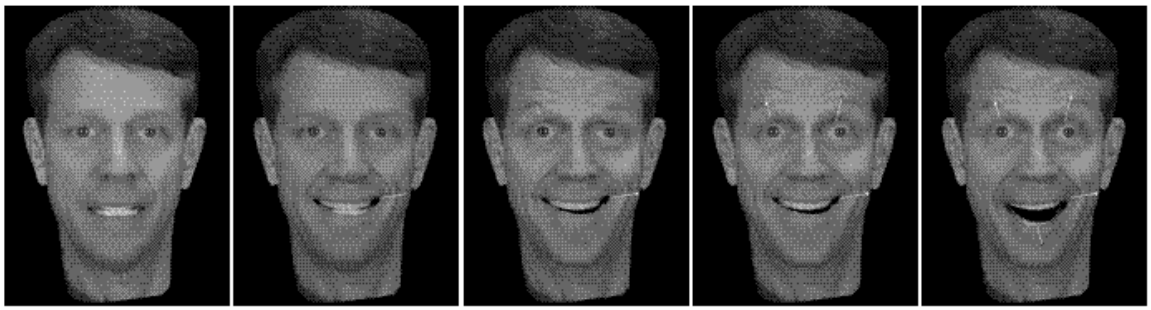


Figura 6-3: Edição sucessiva de um quadro (*keyframe*) a partir do mais grosseiro (imagem mais à esquerda) até o nível de detalhe mais fino (imagem mais à direita).

6.5. RENDERING BLEND SHAPES REALISTAS

O *rendering* do modelo *blend shape* é consideravelmente direto e pode ser feito em dois passos [Joshi2003]: no primeiro passo, a geometria consensual é avaliada, depois ela é renderizada tantas vezes quantos forem os *blend shapes* no modelo para combinar o mapa de textura. Esse último passo é feito através da atribuição do peso correspondente para um dado *blend shape* ao canal *alpha* de cada vértice.

O problema de *texture misregistration* é comum quando se está fazendo o *rendering* do *blend shape* para uma animação facial realista. Se a textura não corresponde em cada ponto à geometria da face, combiná-la linearmente irá gerar um resultado com o *rendering* borrado.

Para aliviar esse problema, o artigo pegou emprestada uma técnica utilizada na comunidade de processamento de imagem: a combinação foi baseada em uma decomposição passa-faixa (*band-pass*) das texturas. Mais especificamente foram construídos dois níveis de pirâmide de imagem laplaciana de cada mapa de textura do *blend shape*. Isso resultou na criação de dois mapas de textura para cada *blend shape*: o primeiro é uma versão passa-baixa da textura original e o segundo é a textura detalhada. O resultado é um *rendering* que preserva melhor o conteúdo original do espectro das texturas *blend shape* e que mantém o conteúdo de alta freqüência constante em toda animação.

6.6. RESULTADOS E TRABALHOS FUTUROS

A técnicas descritas neste artigo foram demonstradas com um conjunto de *blend shapes* modelados que foram capturados a partir da expressão facial de um ator. Foram criados 10 *blend shapes* correspondendo às expressões extremas. Três fotografias do ator foram processadas para modelar cada *blend shape*: uma de frente, uma a 30 graus para a esquerda e outra a 30 graus para a direita.

Como trabalhos futuros o artigo aponta os seguintes tópicos:

- O algoritmo de *rendering* pode ser melhorado através da análise das freqüências principais dos mapas de textura do *blend shape*;
- Seria interessante aplicar as técnicas do artigo em um personagem não-humano, isto porque a segmentação da face poderia ser mais desafiadora e não-intuitiva;

- Seria interessante testar a técnica em um conjunto maior de *blend shapes*; e, por fim,
- Seria interessante estender a técnica de segmentação para levar em consideração as informações de textura.

6.7. CONCLUSÕES PESSOAIS

i. Quem são os autores?

Pushkar Joshi: Aluno de mestrado na Universidade de Berkeley na área de Computação Gráfica. Esse trabalho é o resultado do trabalho final de graduação na USC.

Mathieu Desbrun: Professor associado no Computer Science no CALTECH (*California Institute of Technology*). Principais áreas de pesquisa: animação interativa, animação em multiresolução, malhas irregulares, ambientes imersivos etc.

Frédéric Pighin: Pesquisador senior do Institute for Creative Technologies (University of Southern California) onde ele é o gerente do Computer Animation Lab. Tem como objetivo desenvolver técnicas para sintetizar atores virtuais foto-realistas, tendo como interesse tanto a representação de atores sintéticos para *rendering* e animação, quanto o desenvolvimento de ferramentas para permitir que artistas digitais criem seus personagens humanos com aparência "*believable*".

ii. O que o artigo resolve?

O artigo desenvolveu uma técnica automática para extrair um conjunto de parâmetros de um modelo *blend shape*. A partir desse conjunto é possível combinar as imagens e conseguir desenvolver uma animação. Com a segmentação e os *blend shapes* é possível fazer a animação do personagem tanto através da captura de movimento quanto através da animação de quadros-chave (animação *keyframing*).

iii. Qual a abordagem utilizada?

O artigo assumiu que cada *blend shape* pode ser descrito como uma combinação linear, assumiu também um modelo físico cuja sua elasticidade pode ser descrita através da equação de Lamé e a partir desse modelo conseguiu calcular o valor do laplaciano discreto para cada vértice das expressões não-neutras. A partir desse último cálculo, pegando a magnitude dos vetores resultantes, foi possível obter um mapa de deformação para cada expressão. A partir desse mapa e com o uso de um limiar pode-se obter a segmentação da face em regiões.

iv. Qual a classificação do artigo?

O artigo caracteriza-se por ser de aprendizado de expressões a partir de *blend shapes*. Assim como outros artigos apresentados neste relatório, ele busca uma animação facial fotorealista.

v. Quais foram as ferramentas utilizadas na implementação?

O artigo não comenta a linguagem utilizada no desenvolvimento do sistema apresentado, ele aborda apenas a matemática necessária para a concepção de cada fase. Apenas é comentado que as animação foram computadas e renderizadas em tempo real (30Hz) em um PC de 1GhZ equipado com uma placa gráfica NVidia GeForce 3.

vi. Quais os problemas em aberto interessantes que o artigo apresenta ("ataca")?

O artigo aborda a técnica de aprendizado de expressões para um dado modelo facial. O algoritmo utilizado para a segmentação da face também demonstrou ser bastante eficiente e preciso.

vii. O que não foi feito neste artigo que é interessante? (Quais os problemas em aberto interessantes que o artigo desperta e não ataca?)

Em instante algum do artigo, foi abordado a animação do personagem blend *shape* com fala, apenas as expressões faciais são tratadas. Um ponto interessante seria verificar a possibilidade de fazer o casamento de uma técnica dessa de aprendizado levando em consideração um sistema *talking head*.

7. Artigo VI: "An Example-Based Approach for Facial Expression Cloning"

O artigo [Pyun2003] apresenta uma nova abordagem baseada em exemplos para clonagem de expressões faciais. A partir de um conjunto de exemplos de modelos-chave de uma determinada face origem, a idéia é conseguir uma clonagem das expressões faciais do modelo origem em um modelo destino (face destino) preservando as características faciais deste modelo alvo na animação final.

7.1. INTRODUÇÃO

A síntese através do reuso de dados existentes tem se tornado uma área popular na computação gráfica [Pyun2003]. Inspirados no movimento de *retargetting*, Noh e Neumann [Noh2001] trabalharam no problema de clonagem de expressão facial de um modelo de face 3D existente para um novo modelo. Baseado em *morphing* de geometria 3D, a solução proposta por eles para esse problema foi primeiro computar os vetores de movimento do modelo origem para depois deformar esses vetores e adicionar ao modelo alvo. Essa abordagem funciona bem para modelos de face com formatos similares.

Existe um incentivo na área de pesquisa de animação facial dirigida por parâmetro tais como sistema de codificação de ações faciais ou transmissão de pessoa baseada em um modelo [Parke1996]. Baseado na noção de transferência de parâmetros, Bregler et al. [Bregler2002] propuseram um esquema elegante de captura de movimento e *retargetting*. Eles escolhem exemplo de *key-shapes* de um modelo origem de uma animação *cartoon* dada como entrada e modela os *key-shapes* correspondente para um modelo destino.

O restante desta seção destina-se a apresentar alguns dos trabalhos relacionados a este artigo e a apresentar uma visão geral do trabalho desenvolvido.

7.1.1. Trabalhos Relacionados

Vem existindo um esforço extensivo no desenvolvimento de técnicas de animação facial 3D desde o trabalho pioneiro de Parke [Parke1972]. O trabalho descrito no artigo [Pyun2003] começou lidando com abordagens tradicionais para geração de animação facial como *scratch* (riscar) e depois foi movendo para trabalhos mais recentes diretamente relacionado ao esquema proposto (desejado).

Dentre os trabalhos de "*Facial Animation From Scratch*", Parke propôs uma abordagem paramétrica para representar o movimento de um grupo de vértices com um vetor de parâmetros e usou essa abordagem para gerar um grande conjunto de expressões faciais [Parke1982]. Em abordagens baseadas em performance, animações faciais eram sintetizadas de acordo com o movimento facial de dados capturados a partir da performance de um ator real, como no trabalho [Williams1990]. Pighin et al. [Pighin1998] desenvolveu um trabalho cuja a abordagem era baseada em imagens para gerar expressões faciais 3D fotorealistas a partir de um conjunto de fotografias 2D. Todas as abordagens desenvolvidas tinham em comum o fato de ser necessário repetir todo o

procedimento para animar um novo modelo facial, mesmo que uma seqüência de expressões similar estivesse disponível para um modelo diferente.

Uma outra área de trabalhos relacionados é a de "*Retargetting and Cloning*", que mostra que recentemente vêm existindo resultados ricos de pesquisa na área de reuso de dados existentes de animação [Pyun2003]. Gleicher [Gleicher1998] descreveu um método para *retargetting* de movimento para um novo personagem com diferentes proporções de segmentos. Noh e Neumann [Noh2001] adaptaram a idéia de *retargetting* de movimento para o reuso de dados de animação facial. Baseados no *morphing* de geometria 3D entre os modelos de face origem e destino com o objetivo de gerar expressões clonadas no modelo alvo.

Por fim, o trabalho desenvolvido relaciona-se também com trabalhos da área "E*xample-based Motion Synthesis*". Bregler et al. [Bregler2002] propôs uma abordagem baseada em exemplo para captura de movimento e *retargetting* de *cartoon*. Baseado em transformações afins, a abordagem deles primeiro extrai os parâmetros de transformação e os pesos de interpolação dos formatos-chave origem em cada quadro da animação *cartoon* de entrada. Então essa abordagem gera o formato de saída aplicando tanto os parâmetros quanto os pesos aos formatos-chave do alvo (*target*) correspondente. A abordagem utilizada não é trivial para ser adaptada no problema apresentado neste artigo de clonagem de expressão facial [Joshi2003]. Existem abordagens similares a essa no domínio da animação dirigida por performance para *retargetting* de expressão facial a partir de vídeos 2D para desenhos 2D [Buck2000] ou para modelos 3D [Chuang2002b].

7.1.2. Visão Geral

Inspirados na captura de movimento e no *retargetting* de *cartoon* [Bregler2002], o trabalho deste artigo adaptou uma abordagem baseada em exemplo para *retargetting* expressões faciais a partir de um modelo para outro. A Figura 7-1 ilustra uma visão geral da abordagem baseada em exemplo para clonagem de expressão desenvolvida no artigo [Joshi2003].

Dada como entrada uma animação facial 3D para um modelo origem de face, é possível gerar uma animação similar para um modelo alvo através da combinação de modelos-alvo predefinido correspondentes aos exemplos do modelo origem de face com expressões extremas chamadas de *key-models* (modelos-chave). A abordagem desenvolvida é constituída, basicamente, de três etapas: *model construction* (construção do modelo), *parametrization* (parametrização) e *expression blending* (combinando expressões). As duas primeiras etapas são feitas uma única vez e no início, e a última fase é executada repetidamente para cada expressão de entrada em tempo de execução. Essas três etapas serão explicadas em detalhes, cada uma individualmente como também a correspondência entre elas, nas próximas deste capítulo.

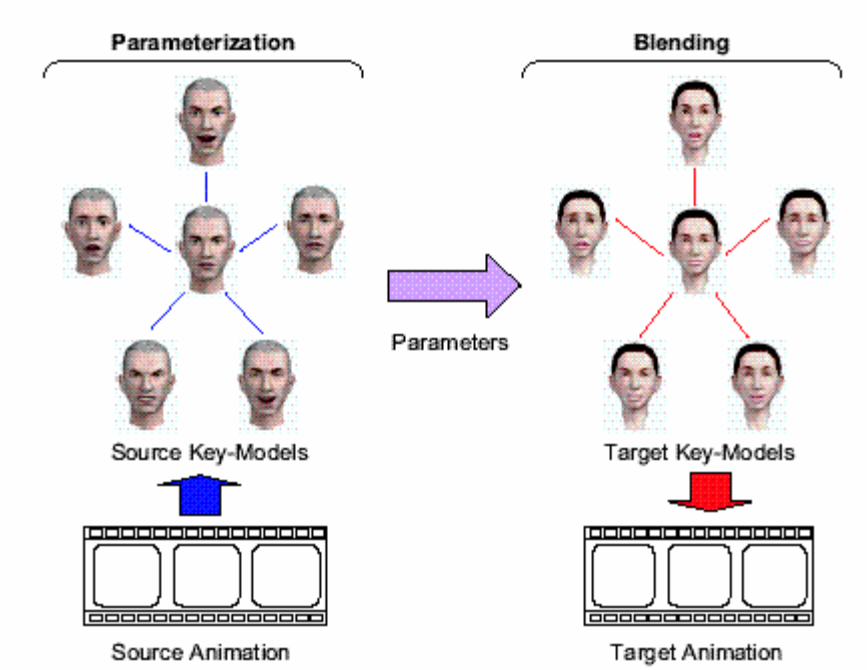


Figura 7-1: Visão geral do sistema de clonagem baseado em exemplo.

7.2. CONSTRUÇÃO DO MODELO-CHAVE

O problema de clonagem de expressões faciais tem uma natureza um pouco diferente do problema de captura de movimento e *retargetting* em *cartoon*. Nesse último, os personagens *cartoon* têm a capacidade de modificar seus formatos dinamicamente e ainda continuar preservando a suas identidades.

Por outro lado, expressões faciais são determinadas através da combinação sutil de deformações locais em um modelo origem de face ao invés de ser a mudança global de uma forma. Diferente dos movimentos de *cartoon*, movimentos de face (expressões) devem ser bem caracterizadas.

Em particular, o trabalho desenvolvido neste artigo, definiu duas categorias de expressões-chave: expressões-chave emocionais e expressões-chave verbais. A primeira categoria de expressões-chave reflete estados emocionais, enquanto que a segunda reflete resultados de movimentos labiais devido à comunicação verbal (a fala). O artigo combina essas duas categorias para definir expressões-chave genéricas de um modelo origem e depois cria os modelos-chave correspondentes tanto para os modelos de face origem quanto alvo (destino ou *target*) através da deformação dos seus modelos de base respectivos com a expressão neutra.

Como ilustra a Figura 7-2, foram escolhidas seis expressões-chave emocionais puras: *neutral* (neutra), *happy* (feliz), *sad* (triste), *surprised* (surpresa), *afraid* (com medo) e *angry* (com raiva). A expressão neutra foi escolhida como expressão base. Baseada na noção de visema, foram escolhidos 13 visemas como as expressões-chave puramente verbais, como ilustra Figura 7-3. Combinando essas duas categorias de expressões-chave tem-se 78 (6 * 13) expressões-chave genéricas junto com seis expressões puramente emocionais.

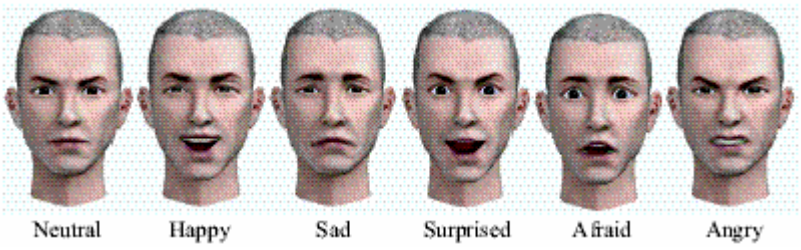


Figura 7-2: As seis expressões-chave emocionais. Da esquerda para direita tem-se: neutra, feliz, triste, surpresa, com medo e com raiva.

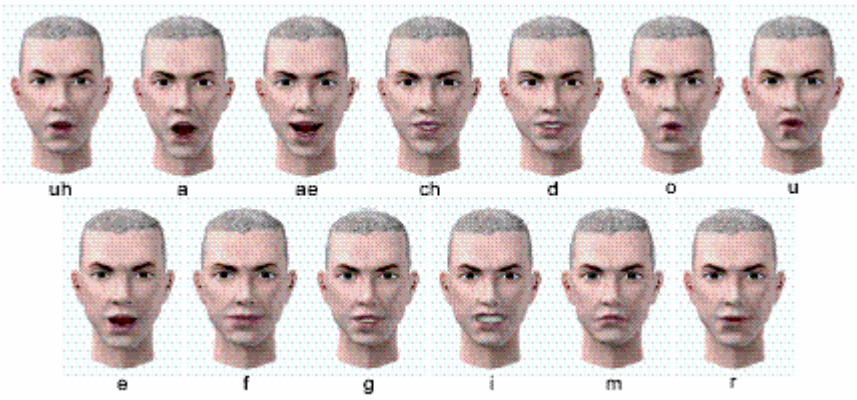


Figura 7-3: As treze expressões verbais.

Para facilitar a clonagem de expressão baseada em exemplo, foi necessário ter 84 modelos-chave correspondentes para cada um dos modelos de face origem e destino. Agora o problema de automatizar a criação dos modelos-chave foi reduzido a um problema de composição de geometria: dada uma expressão-chave emocional, como é possível obter suas expressões-chave combinadas?

Sem perda de generalidade, foi necessário supor que os modelos de face eram representados por malhas poligonais (poliedros). Então, as expressões causam movimentos dos vértices no modelo da face. Analisando os modelos-chave emocional e verbal, os vértices da malha foram caracterizados segundo suas contribuições para as expressões faciais. Por exemplo, vértices próximos aos olhos contribuem principalmente para a emoção, enquanto que vértices próximos à boca contribuem tanto para expressões emocionais quanto verbais.

Baseado nessas contribuições sobrepostas, o artigo introduziu a noção de importância que mede a contribuição relativa de cada vértice para as expressões verbais levando em consideração as expressões emocionais. O valor da importância α_i de cada vértice v_i é estimado empiricamente através dos modelos-chave verbal e emocional de tal modo que $0 \le \alpha_i \le 1, \forall i$. Se $\alpha_i \ge 0.5$ então o movimento do vértice v_i é restrito às expressões verbais; caso contrário, é restrito às expressões emocionais. A Figura 7-4 ilustra a distribuição da importância dos valores sobre um modelo de face e o apêndice do

artigo [Joshi2003] apresenta um detalhamento dessa distribuição para o esquema facial proposto por eles.

Figura 7-4: Distribuição da importância dos valores.

Agora vem o ponto de explicar como o artigo fez para compor um modelo-chave emocional *E* e um modelo-chave verbal *P* derivados a partir de um modelo de base *B*. Sejam $B = \{v_1, v_2, ..., v_n\}$, $E = \{v_1^E, v_2^E, ..., v_n^E\}$ e $P = \{v_1^P, v_2^P, ..., v_n^P\}$. Tem-se que v_i^E e v_i^P, com $1 \le i \le n$ são obtidos pelo deslocamento de v_i, se necessário. Para cada vértice v_i, foram definidos os deslocamentos Δv_i^E e Δv_i^P como sendo $\Delta v_i^E = v_i^E - v_i$ e $\Delta v_i^P = v_i^P - v_i$, respectivamente.

Sejam $C = \{v_1^C, v_2^C, ..., v_n^C\}$ e $\Delta v_i^C = v_i^C - v_i$, respectivamente, o modelo-chave combinado e o deslocamento de um vértice $v_i^C \in C$. Como o problema de computar v_i^C pode ser reduzido a composição de dois vetores Δv_i^E e Δv_i^P, assumiu-se que v_i^C cai (pertence) ao plano expandido por Δv_i^E e Δv_i^P e contendo v_i, como ilustra a Figura 7-5.

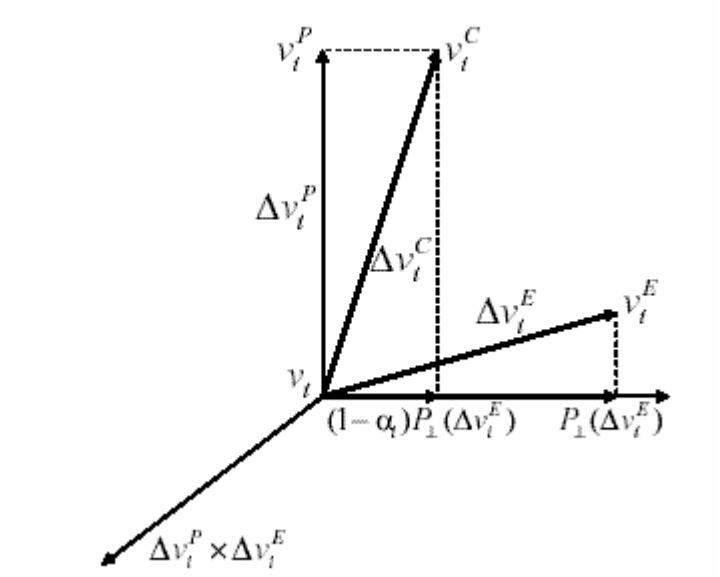


Figura 7-5: Composição de dois deslocamentos.

Considerando o vértice v_i^C, com $1 \le i \le n$, pertencente ao modelo combinado *C*. Se $\alpha_i \ge 0.5$ então o componente verbal Δv_i^P deve ser preservado em Δv_i^C para garantir a pronúncia correta. Conseqüentemente, seja $P \perp \left(\Delta v_i^E\right)$ o componente de Δv_i^E perpendicular a Δv_i^P, apenas esse componente $P \perp \left(\Delta v_i^E\right)$ de Δv_i^E pode contribuir para Δv_i^C no topo de Δv_i^P. Se $\alpha_i < 0.5$ então as regras para Δv_i^E e Δv_i^P são trocadas.

Resumidamente, tem-se:

$v_i^C = \begin{cases} v_i + \left(\Delta v_i^P + (1-\alpha_i) P \perp \left(\Delta v_i^E\right)\right) \\ v_i + \left(\Delta v_i^E + \alpha_i E \perp \left(\Delta v_i^P\right)\right) \end{cases}$, no primeiro caso se $\alpha_i \ge 0.5$ e no segundo, caso contrário;

onde $P \perp \left(\Delta v_i^E\right) = \Delta v_i^E - \frac{\Delta v_i^E \cdot \Delta v_i^P}{\left|\Delta v_i^P\right|^2} \cdot \Delta v_i^P$ e $E \perp \left(\Delta v_i^P\right) = \Delta v_i^P - \frac{\Delta v_i^P \cdot \Delta v_i^E}{\left|\Delta v_i^E\right|^2} \cdot \Delta v_i^E$

A Figura 7-6 ilustra um modelo chave combinado (Figura 7-6 (c)) construído a partir de um modelo-chave com uma expressão verbal (Figura 7-6 (a)) e um modelo-chave com expressão emocional (Figura 7-6 (b)). É importante observar que a expressão verbal é preservada ao redor da boca e a expressão emocional nas outras partes.

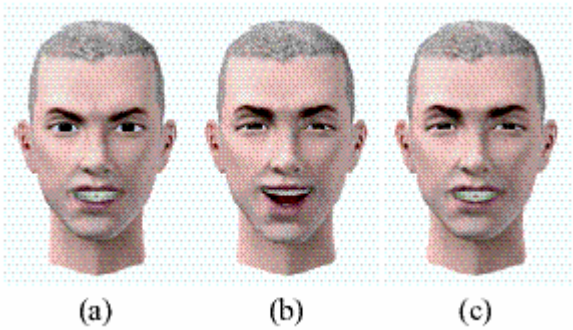


Figura 7-6: Composição de modelos-chave: (a) modelo-chave verbal (vogal "i"); (b) modelo-chave emocional ("*happy*"); e (c) modelo-chave combinado.

7.3. PARAMETRIZAÇÃO

A etapa de parametrização dos modelos-chave destino foi feita baseada na correspondência entre o modelo de base origem e os modelos-chave origem. Interativamente foi selecionado um número de pontos característicos do modelo de base origem e depois foi extraído os seus deslocamentos para os pontos correspondentes em cada um dos modelos-chave origem. Concatenando esses deslocamentos, o vetor deslocamento de cada modelo-chave origem é formado para parametrizar o modelo-chave destino correspondente. Assim, baseado em PCA (*Principal Components Analysis* ou Análise de Componentes Principais) [Jollife1986], a dimensionalidade do espaço de um parâmetro pode ser reduzida através da remoção dos vetores de base menos

significativos do *eigenspace* resultante. Como ilustra a Figura 7-7 foram selecionados cerca de 20 pontos característicos da base do modelo origem.

Figura 7-7: Base do modelo-chave origem com 20 pontos característicos selecionados manualmente.

O vetor deslocamento v_i de um modelo-chave origem S_i a partir da base de um modelo-chave origem S_B é definido por $v_i = s_i - s_B, 1 \le i \le M$, onde s_B e s_i são vetores obtidos pela concatenação, em uma ordem fixa, de coordenadas 3D de pontos característicos em S_B e alguns em S_i, respectivamente, e M é o número de modelos-chave de origem. Como ilustra a Figura 7-8, as posições de v_i de cada modelo-chave destino T_i em um espaço de parâmetros N-dimensional, onde N é o número de componentes, isto é, três vezes o número de pontos característicos.

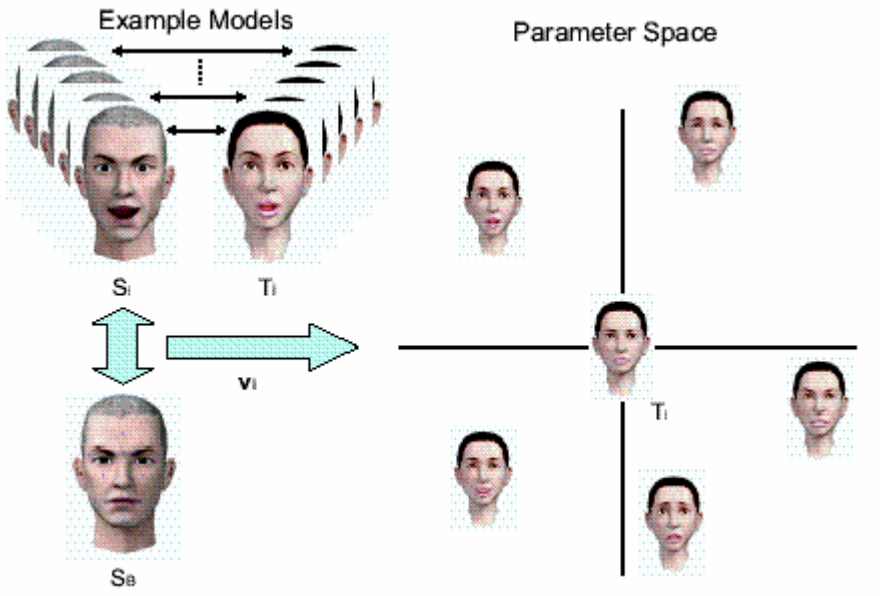


Figura 7-8: O vetor deslocamento de cada modelo-chave origem S_i é usado para a parametrização do modelo-chave destino correspondente T_i.

Seja e_i, $1 \le i \le N$ o autovetor correspondente ao i-ésimo maior autovalor. Supondo que são escolhidos $\overline{N}$ autovetores como os eixos de coordenadas do espaço de parâmetro, onde $\overline{N} < N$. Para transformar um vetor deslocamento original N-

dimensional em um vetor de parâmetro $\overline{N}$-dimensional, uma matriz ***F*** de dimensão $\overline{N} \times N$ é construída. Essa matriz ***F*** é chamada de matriz característica: $F = [e_1 e_2 e_3 ... e_{\overline{N}}]^T$.

Usando a matriz característica ***F***, o vetor de parâmetro p_i correspondente ao vetor de deslocamento v_i do modelo-chave destino, T_i é derivado a partir de $p_i = Fv_i, 1 \le i \le M$, e assim reduzindo a dimensionalidade do espaço de parâmetro de N para $\overline{N}$. A matriz ***F*** será utilizada posteriormente para computar o vetor de parâmetro a partir de um dado vetor deslocamento.

7.4. COMBINANDO EXPRESSÕES

Com os modelos-chave parametrizados, o problema de clonagem foi transformado em um problema de interpolação de dados espalhados (*scattered*). Para solucionar esse problema, o esquema de combinar as expressões predefine funções de peso para cada modelo-chave destino baseado-se em funções de base cardinal que consiste de funções de base linear e funções de base radial.

O formato global de uma função de peso é primeiro aproximado por uma função de base linear e depois ajustado localmente por funções de base radial para interpolar exatamente o modelo-chave correspondente. Dado um modelo de face como entrada com uma expressão facial, um novo modelo de saída é gerado com a expressão clonada, e é obtido em tempo de execução através da combinação dos modelos-chave destino como ilustra a Figura 7-9.

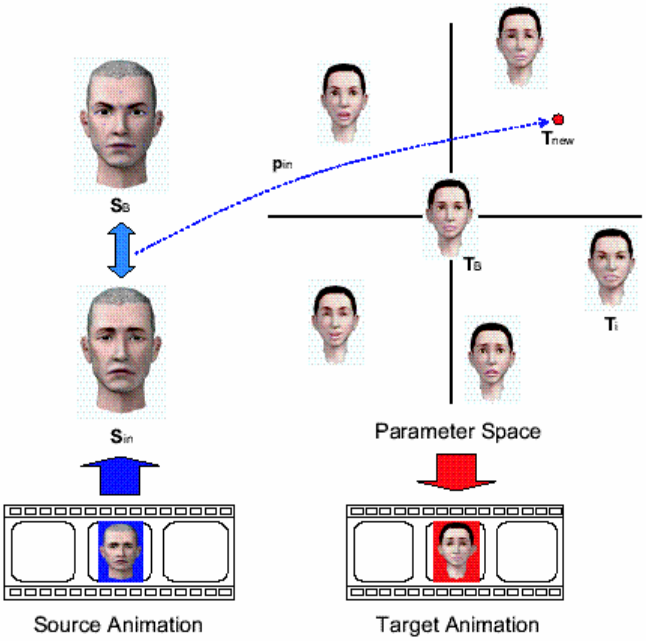


Figura 7-9: Geração de um novo modelo de face através da combinação de modelos-chave destino.

A função de peso $w_i(\cdot)$ de cada exemplo de modelo destino T_i, $1 \le i \le M$ em cada parâmetro do vetor p é definido por $w_i(p) = \sum_{i=0}^{\overline{N}} a_{il} A_l(p) + \sum_{j=1}^{M} r_{ji} R_j(p)$ [Pyun2003]. Com as funções de peso predefinidas, é possível verificar como combinar modelos-chave destino em tempo de execução.

Para um modelo de face de entrada S_{in} em cada quadro da animação de entrada o vetor deslocamento d_{in} é computado com relação ao modelo de base origem S_B: $d_{in} = s_{in} - s_B$, onde s_i e s_B são, respectivamente, vetores obtidos através da concatenação de coordenadas 3D dos pontos característicos em S_{in} e S_B. Dado esse vetor deslocamento N-dimensional d_{in}, é possível obter o vetor de parâmetro correspondente $\overline{N}$-dimensional p_{in} através de $p_{in} = Fd_{in}$, onde $\boldsymbol{F}$ é a matriz de característica.

Usando as funções de peso predefinidas para os modelos-chave destino T_i, é possível estimar os valores dos pesos $w_i(p_{in})$ de todos os modelos-chave destino T_i, $1 \le i \le M$ no parâmetro p_{in} para gerar o modelo de face de saída $T_{new}(p_{in})$: $T_{new}(p_{in}) = T_B + \sum_{i=1}^{M} w_i(p_{in})(T_i - T_B)$, onde T_B é o modelo base destino correspondente ao modelo-chave base origem S_B com a expressão neutra.

7.5. RESULTADOS EXPERIMENTAIS

Como ilustra a Figura 7-10, foram utilizados dois modelos origem e quatro modelos destino nos experimentos desenvolvidos no artigo. A tabela ilustrada na Figura 7-11 mostra o número de vértices (V) e o número de polígonos (P) em cada modelo. Como animações de entrada foram preparadas duas diferentes animações: uma animação facial do "*Man A*" com várias expressões exageradas e uma animação facial do "*Man B*" com expressões verbais combinadas com expressões emocionais.

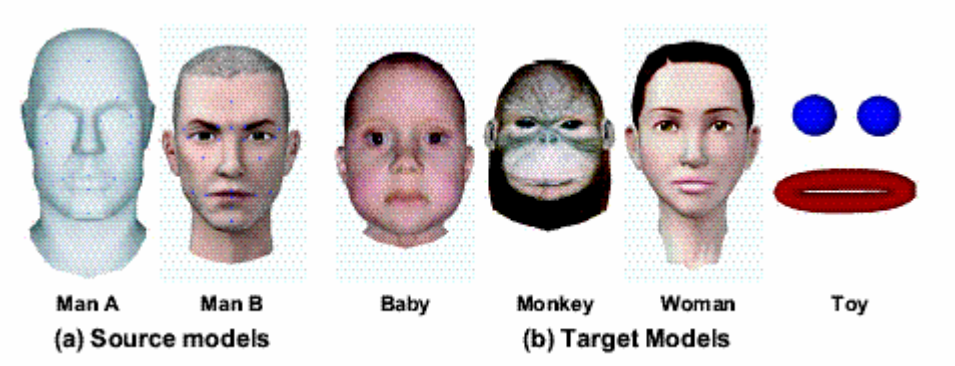


Figura 7-10: Modelos utilizados nos experimentos.

	Man A	Man B	Baby	Monkey	Woman	Toy
V	988	1192	1253	1227	1220	931
P	1954	2194	2300	2344	2246	957

Figura 7-11: Especificação do modelo.

No primeiro experimento desenvolvido, foi usado o "*Man A*" como modelo de face origem e os modelos do bebê e do macaco como modelos de face destino. Para clonar a expressão facial do "*Man A*" foram utilizados seis modelos-chave para o modelo de face origem e o mesmo para o modelo de face destino. A primeira linha da Figura 7-12 mostra as expressões de entrada do "*Man A*" amostradas a partir da animação de entrada. Os modelos do bebê e do macaco com as expressões faciais clonadas são mostrados, respectivamente, na segunda e na terceira linhas da Figura 7-12.

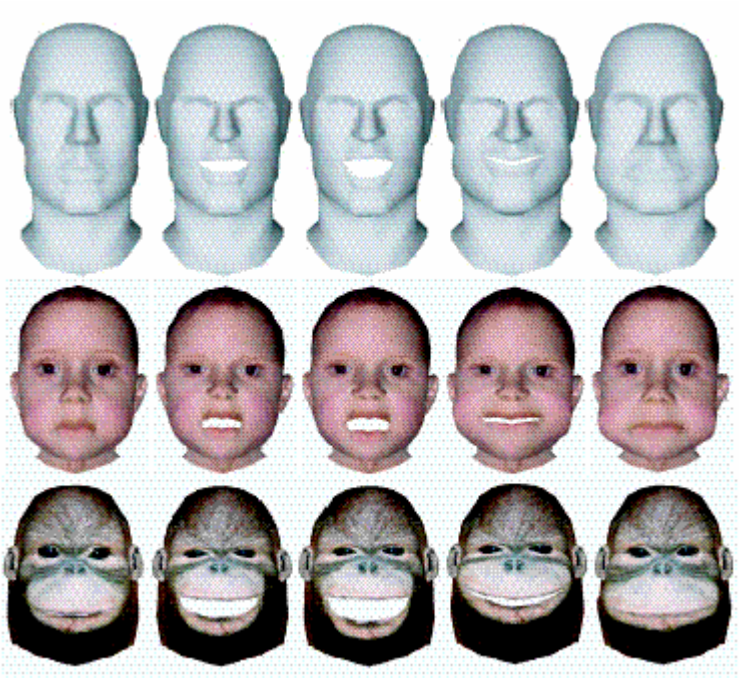


Figura 7-12: Expressões clonadas do *"Man A"* para os modelos destino.

No segundo experimento foi utilizado o "*Man B*" como modelo origem e o modelo da mulher como modelo destino para clonar as expressões combinadas. Foi preparado um total de 84 modelos-chave: seis expressões-chave que são as expressões puras (*neutral* (neutra)*, happy* (feliz)*, sad* (triste)*, angry* (com raiva)*, surprised* (surpresa) e *afraid* (com medo)) e treze visemas (sete para vogais e seis para consoantes) para cada uma das seis expressões-chave emocionais.

Como mostra a Figura 7-13, as expressões combinadas do modelo origem foram convincentemente reproduzidas no modelo destino. Foram também clonadas expressões emocionais do "*Man B*" no modelo topologicamente diferente, como ilustra a Figura 7-14.

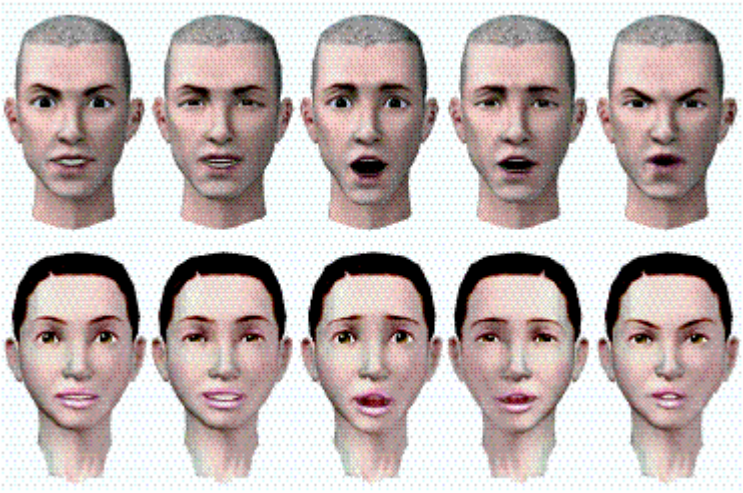


Figura 7-13: Expressões clonadas do "*Man B*" para o modelo destino.

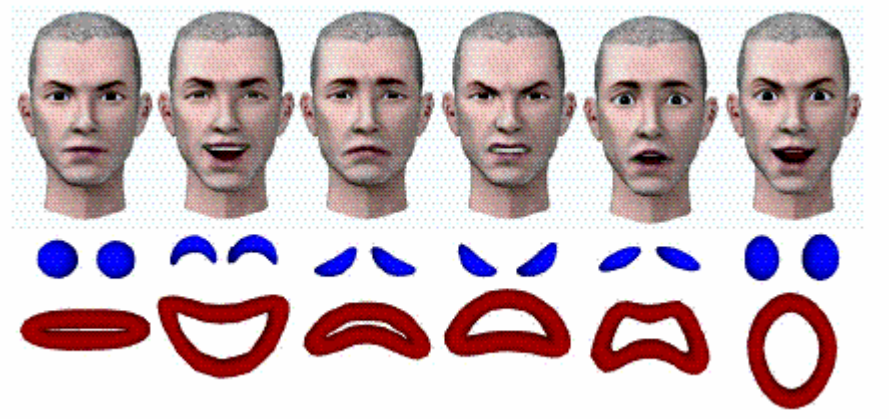


Figura 7-14: Expressões clonadas do "*Man B*" para o modelo topologicamente diferente.

7.6. CONCLUSÕES DO ARTIGO

O artigo [Pyun2003] apresentou uma abordagem inédita para clonagem de expressões faciais baseado em exemplos a partir de um modelo origem para um modelo destino preservando as características do modelo destino. Além dessa principal contribuição do artigo, uma das etapas desse desenvolvimento consistiu de uma contribuição: a etapa de construção do modelo-chave. Para realização dessa fase foi desenvolvido um novo esquema para composição do par de modelos-chave verbal e

emocional. Como mostrado nos resultados experimentais e afirmado no artigo, a abordagem para clonagem de expressões faciais apresentadas tem um resultado eficiente.

Uma limitação da sua abordagem que o artigo apresenta é o fato de necessitar que o animador prepare um conjunto de modelos-chave para os modelos origem e destino como fase do processamento. Uma segunda limitação também apontada é que o método desenvolvido não consegue clonar corretamente uma expressão quando ela se localiza muito fora da base construída do modelo-chave origem.

Como trabalho futuro, o grupo está planejando estender a abordagem apresentada para clonagem de expressões por região. De acordo com resultados da psicologia, a face pode ser dividida em várias regiões que têm unidades coerentes. Por exemplo, os olhos, as sobrancelhas e a testa seriam regiões de expressões emocionais, enquanto que a boca, as bochechas e o queixo seriam usados para expressões verbais.

7.7. CONCLUSÕES PESSOAIS

i. Quem são os autores?

Hyewon Pyun: Membro do *Theory of Computation Lab*, *Computer Science Department*, KAIST.

Yejin Kim: Membro do *Theory of Computation Lab*, *Computer Science Department*, KAIST. Áreas de interesse: nível de detalhe, multiresolução, animação facial (destaca como principal área no momento) e coloca como subáreas de animação facial "animação facial dirigida por emoção" e "clonagem de expressão".

Hyung Woo Kang: (não encontrei página pessoal)

Sung Yong Shin: Professor do *Department of Computer Science* do *Korea Advanced Institute of Science Technology*. Área de interesse: animação por computador, *real-time rendering* e geometria computacional.

ii. O que o artigo resolve?

O artigo apresenta uma técnica interessante para clonagem de expressões faciais. Basicamente, tem-se um modelo de origem e um ou mais modelos de destino (alvo). Considerando uma certa expressão facial feita no modelo facial de origem, é possível clonar essa expressão no modelo destino. O ponto interessante é que não se trata de uma "cópia" de expressões e sim de uma clonagem: não se coloca o personagem destino na mesma expressão do personagem origem, os movimentos faciais são clonados preservando a estrutura facial e a geometria do modelo alvo.

iii. Qual a abordagem utilizada?

Buscou-se utilizar técnicas tradicionais para solução do problema, investigando sempre qual seria a melhor solução e assim, conseqüentemente, chegando até a mesma.

iv. Qual a classificação do artigo?

O artigo classifica-se por ser de clonagem de expressões faciais, tanto para personagens realistas quanto para personagens de estilo *cartoon*.

v. Quais foram as ferramentas utilizadas na implementação?

O sistema de clonagem de expressões faciais foi implementado na linguagem de programação C++ e utilizando OpenGL. A máquina de desenvolvimento e testes consistiu de um PC com processador Intel Pentium 4 com 2.4 GHz, 512 MB de RAM e uma placa gráfica GeForce4.

vi. Quais os problemas em aberto interessantes que o artigo apresenta ("ataca")?

Além do assunto principal de clonagem de expressões faciais que é por si só bastante interessante, o artigo preocupa-se com a questão se um determinado vértice da malha facial é mais um vértice emocional ou um vértice verbal. Isso é um ponto importante, e até comparado aos demais já apresentados, este trabalho é um dos poucos que se preocupa com esse casamento da fala com as expressões faciais.

vii. O que não foi feito neste artigo que é interessante? (Quais os problemas em aberto interessantes que o artigo desperta e não ataca?)

Apesar do artigo ter definido "a importância de um vértice" ele não fez o casamento das importâncias caso um determinado trecho de fala tenha emoção. A diferença que o artigo apontou é que se um vértice tiver uma determinada importância prevalece a expressão facial naquele vértice, caso contrário, é aplicado o visema. O artigo não combina para um mesmo vértice a expressão verbal e emocional.

8. Artigo VII: "Vision-based Control for 3D Facial Animation"

O artigo [Chai2003] apresenta uma ferramenta desenvolvida em que um usuário qualquer faz uma expressão em frente a uma única câmera de vídeo e essa expressão é reproduzida por um avatar. O ponto interessante é que esse mesmo avatar é capaz de reproduzir as expressões de diferentes usuários em tempo real. O artigo é bastante completo e traz contribuições importantes para diversas áreas da computação gráfica e mais particularmente para a animação facial.

8.1. INTRODUÇÃO

Fornecer um controle intuitivo sobre expressões faciais tridimensionais é um problema importante no contexto de áreas como computação gráfica, realidade virtual e interação humano-computador. Duas dificuldades existem nesse campo: desenvolver um conjunto rico de ações convincentes para o personagem virtual, e dar ao usuário o controle sobre essas ações.

Este artigo [Chai2003] tem o propósito de combinar as forças de uma interface baseada em visão com as forças da captura de movimento com o objetivo de fornecer um controle interativo de animações faciais 3D. O artigo vai mostrar que um rico conjunto de ações faciais humanas pode ser criado a partir de uma base de dados de captura de movimento e o usuário pode controlar essas ações interativamente fazendo os movimentos desejados em frente de uma câmera de vídeo, como ilustra a Figura 8-1.

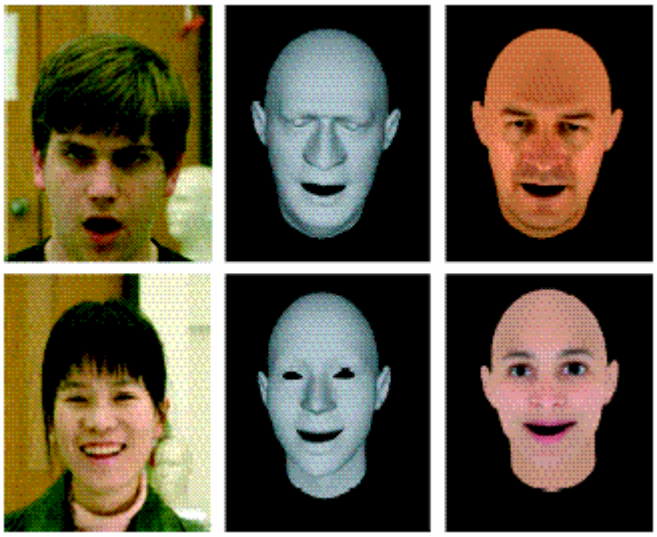


Figura 8-1: Controle Interativo de Expressão: um usuário pode controlar expressões faciais 3D de um avatar interativamente. Na imagem mais à esquerda tem-se o usuário atuando em frente a uma câmera; na imagem ao centro tem-se o movimento facial controlado de um avatar com máscara cinza; e na imagem mais à direita tem-se o movimento facial controlado de um avatar com mapeamento de textura.

No processo desenvolvido, o artigo apresenta como desafios os seguintes tópicos:

- O mapeamento de sinais visuais de baixa qualidade (capacidade da câmera) em movimentos de dados de alta qualidade (ações do ser humano);
- Com o objetivo de controlar as ações faciais interativamente através de uma interface baseada em visão, é necessário extrair sinais significativos de controle da animação da seqüência de vídeo de uma pessoa atuando em tempo real;
- É desejado animar um modelo 3D de personagem através do reuso e da interpolação da captura de movimento de dados, mas a captura de movimento grava apenas o movimento de um número finito de marcas faciais no modelo origem. Assim se faz necessário adaptar a captura de movimento dos dados de um modelo origem a todos os vértices do modelo do personagem a ser animado; e
- Se a idéia do sistema é estar disponível para qualquer usuário controlar o modelo de face 3D, deve-se corrigir as diferenças entre os usuários porque cada pessoa tem diferentes proporções e geometria facial.

8.1.1. Trabalhos Relacionados

Como visto no capítulo introdutório desta monografia (Seção 1.2), a animação facial pode ser alcançada através da "interpolação *keyframe*", da "parametrização direta", da "performance do usuário", em "modelos baseados em pseudo-músculos", em "simulação baseada em músculos", e ainda adicionalmente segundo o artigo [Chai2003], em "dados faciais 2D para a fala" e no "movimento completo de dados capturados 3D". Entre essas abordagens, o trabalho apresentado neste artigo tem uma maior relação com a abordagem dirigida por performance e a captura de movimento. O artigo aponta a referência [Parke1996] como um excelente *survey* de todas as abordagens para animação facial.

Um número grande de pesquisadores descreveu e continua descrevendo técnicas para gravar movimentos faciais diretamente de vídeo. Por exemplo, Terzopoulos e Waters [Terzopoulos1993] rastrearam (*tracked*) contornos característicos das sobrancelhas e dos lábios para uma estimativa automática dos parâmetros de contração dos músculos faciais a partir de uma seqüência de vídeo e esses parâmetros musculares foram utilizados para animar fisicamente uma estrutura baseada em músculos de um personagem sintético. Mais recentemente, Gokturk et al. [Gokturk2001] aplicaram PCA (análise de componentes principais) no rastreamento (*tracking*) de dados estéreos para o aprendizado de um modelo deformável e depois eles incorporaram esse "modelo deformável ensinado" em um *framework* de estimativa de fluxo óptico para, simultaneamente, rastrear a cabeça e um pequeno conjunto de características faciais.

Para o uso direto de movimento rastreado para animação é necessário que o modelo da face do ser humano (pessoa a ser capturada ou rastreada) tenha proporções e geometria similares as do modelo a ser animado. Recentemente, a abordagem de rastreamento baseado em visão combinou técnicas de interpolação de *blendshape* [Lewis2000] a criação de animações faciais para um novo modelo destino

(alvo)[6][Buck2000][Chuang2002b]. O rastreamento baseado em visão pode extrair os valores das expressões faciais em uma imagem de vídeo e os exemplos podem ser interpolados apropriadamente.

Assim como a animação baseada em visão, captura de movimento também faz uso de medidas de dados de movimentos humanos para animar expressões faciais. Captura de dados de movimento possui mais detalhes sutis do que dados de rastreamento de visão porque a exatidão do *setup* do *hardware* é usada em um ambiente capturado controlado. Guenter et al. [Guenter1998] criaram um sistema impressionante para captura de expressões faciais usando marcas faciais especiais e múltiplas câmeras calibradas e reapresentaram essas expressões em um *talking head* 3D altamente realista. Mais recentemente, Noh e Neumann [Noh2001] apresentaram uma técnica para clonagem de expressão que adapta dados de movimento capturado existente de um modelo facial 3D de origem em um novo modelo destino (alvo) 3D. Um exemplo mais recente ainda de captura de movimento de dados ocorreu no filme "*The Lord of the Rings: the Two Towers*" ("O Senhor dos Anéis: As Duas Torres") onde movimentos e expressões faciais pré-gravadas foram utilizadas para animar o personagem sintético "Gollum".

Ambas as abordagens, baseada em visão e captura de movimento, possuem vantagens e desvantagens. A abordagem baseada em visão fornece um caminho intuitivo e "barato" para controlar um grande número de ações, mas seus resultados atuais têm sido desapontadores para aplicações de animação [Chai2003]. Por outro lado, a captura de movimento de dados gera uma animação de alta qualidade mas "cara" para coletar. E uma vez coletada, pode não ser exatamente o que o animador precisa, principalmente para aplicações interativas. O trabalho desenvolvido neste artigo buscou usar as vantagens de cada uma das abordagens e, conseqüentemente, evitar as suas desvantagens. Em particular, foi utilizada uma interface baseada em visão para extrair um pequeno conjunto de parâmetros de controle da animação a partir de uma única câmera de vídeo e então utilizar o conhecimento de captura de movimento dos dados para transladar os dados capturados em expressões faciais de alta qualidade. O resultado é uma animação que faz o que o animador deseja que faça, mas tendo a mesma alta qualidade da captura de movimento.

8.1.2. Visão Geral

O sistema desenvolvido transforma movimento rastreado de baixa qualidade em animação de alta qualidade usando o conhecimento de movimento facial humano que está embutido no movimento capturado dos dados. A entrada do sistema consiste de um único fluxo de vídeo que tenha gravado movimentos faciais de um usuário, uma base de dados de movimentos capturados pré-processados, uma superfície 3D como origem escaneada a partir do movimento capturado de uma pessoa e um modelo de superfície de um avatar 3D a ser animado.

Através da ação da expressão desejada em frente à câmera, um usuário tem um controle interativo sobre as expressões faciais do modelo de personagem 3D. O sistema está organizado em quatro principais componentes: *video analysis* (análise do vídeo), *motion capture data preprocessing* (pré-processamento dos dados da captura de

[6] FaceStation. URL: http://www.eyematic.com

movimento), *expression control and animation* (controle da expressão e da animação) e *expression retargetting* (re-alvo das expressões), como ilustra Figura 8-2. Cada uma dessas etapas será explicada em detalhes ao longo das próximas seções deste capítulo.

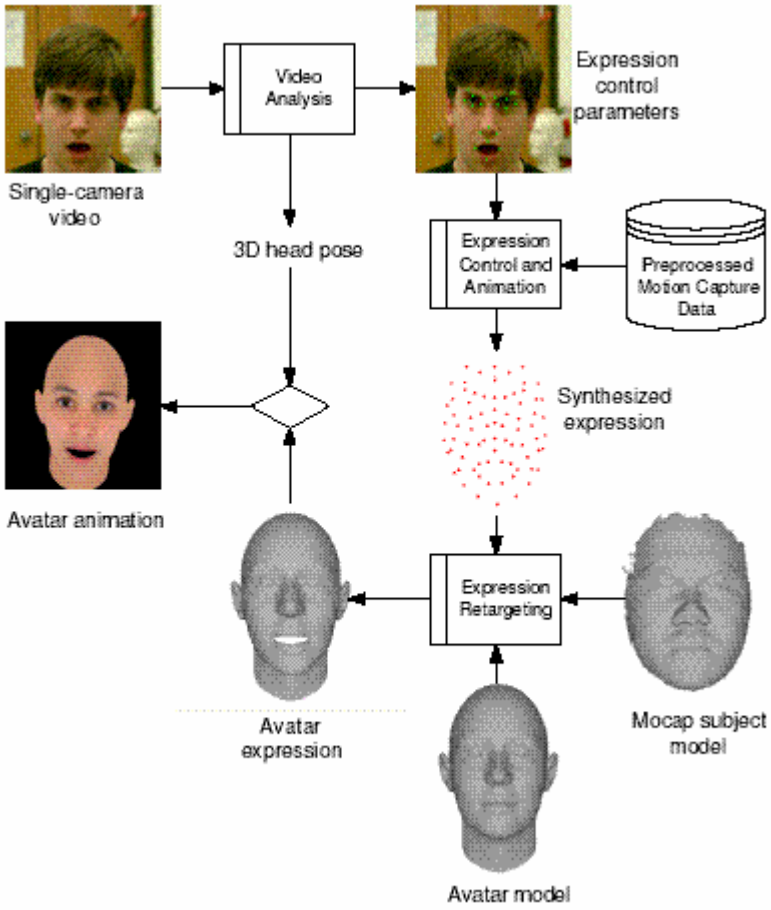


Figura 8-2: Diagrama de visão geral do sistema. Em tempo de execução, imagens de vídeo de uma única câmera são capturadas pelo componente *Video Analysis* que automaticamente extrai dois tipos de parâmetros de controle da animação: parâmetros de controle da expressão e parâmetros de controle da pose 3D. O componente *Expression Control and Animation* usa os parâmetros de controle da expressão e a base de dados de movimentos capturados pré-processados para sintetizar a expressão facial, esta última descrevendo apenas o movimento das marcas da captura de movimento na superfície da pessoa capturada. O componente *Expression Retargetting* utiliza a expressão sintetizada, em conjunto com modelo de superfícies escaneado da pessoa capturada e o modelos de superfície do avatar dado como entrada, para produzir a expressão facial para o avatar. A expressão do avatar é então combinada com a pose do avatar, que é diretamente derivada dos parâmetros de controle da pose, para gerar a animação final.

8.2. ANÁLISE DO VÍDEO

Um sistema de animação baseado em visão considerado ideal deve fazer o rastreamento (*tracking*) tanto do movimento 3D da cabeça quanto das deformações de uma face humana em imagens de vídeo com exatidão. Existem vários algoritmos de rastreamento facial baseado em visão na literatura de visão computacional, mas o

desempenho desses algoritmos não é muito bom [Chai2003]. Conseqüentemente, o trabalho desenvolvido neste artigo não se deteve em rastrear todos os detalhes de uma expressão facial. Eles optaram em, de forma robusta, rastrear um conjunto pequeno de características faciais distintas em tempo real e depois transladar esses dados capturados de baixa qualidade em animação de alta qualidade usando a informação contida em uma base de dados da captura de movimento.

8.2.1. Rastreamento Facial

O sistema desenvolvido no artigo rastreia seis DOFs do movimento da cabeça (*yaw, pitch, roll* e as posições 3D). Foi utilizado um modelo de um cilindro genérico para aproximar a geometria da cabeça do usuário e depois foi aplicada uma técnica de rastreamento baseada no modelo para gravar as posições da cabeça do usuário em um fluxo monocular de vídeo.

O passo de rastreamento de uma expressão resultou em rastrear 19 características 2D da face: um para cada ponto central dos lábios superior e inferior, um para cada canto da boca, dois para cada sobrancelha, quatro para cada olho e três para o nariz, como ilustra a Figura 8-3. A razão de escolher esses pontos característicos é porque os mesmos possuem um alto contraste de textura e porque eles gravam os movimentos de importantes áreas faciais.

Figura 8-3: Rastreamento facial independente do usuário: as setas vermelhas indicam a posição e a orientação da cabeça e os pontos verdes mostram as posições dos pontos de rastreamento (*tracking*).

O algoritmo de rastreamento facial é baseado na inicialização manual do primeiro quadro. Por definição, o sistema inicia com a expressão natural na posição em frente à câmera. O usuário deve clicar nas posições 2D dos 19 pontos no primeiro quadro. Depois que os pontos característicos são identificados no primeiro quadro, o sistema automaticamente computa os parâmetros do modelo cilíndrico. Depois do passo de inicialização, o sistema constrói um mapeamento de textura tendo como referência o modelo da cabeça para ser usado na projeção da primeira imagem na superfície do modelo cilíndrico inicializado.

Devido ao fato que o modelo de referência da cabeça é dinamicamente atualizado, erros de rastreamento podem ser acumulados ao longo do tempo. Foi utilizada uma técnica de re-registro para prevenir esse problema. Em tempo de execução, o sistema automaticamente armazena vários modelos da cabeça com a textura mapeada em poses chaves e escolhe a opção de registrar uma nova imagem de vídeo com o exemplo mais próximo para o modelo de referência da cabeça.

Depois da poses da cabeça terem sido gravadas, o sistema utiliza essas poses e o modelo da cabeça para fazer o *warp* das imagens em uma visão paralela-frontal. As posições das características faciais são então estimadas na imagem atual *warped.*

Para rastrear uma posição 2D de uma característica facial, foi definida uma pequena janela quadrada centralizada na posição característica. Foi feita uma suposição para simplificação que o movimento dos *pixels* na janela característica pode ser aproximado como um movimento afim de um plano centralizado na coordenada 3D da característica. Deformações afim possuem 6 graus de liberdade e podem ser inferidas utilizando o fluxo óptico na janela característica. Foi aplicado um método de estimação do movimento baseado no gradiente para encontrar os parâmetros do movimento afim e com isso conseguiu-se minimizar a soma da diferença de intensidade quadrática na janela característica entre o quadro atual e o quadro de referência [Baker2003].

O sistema de rastreamento facial desenvolvido no artigo executa 20 quadros/segundo em tempo real. O sistema é independente de usuário e pode rastrear o movimento facial de diferentes pessoas, como ilustra a Figura 8-3 onde ele faz o rastreamento da pose e da expressão de duas pessoas.

8.2.2. Parâmetros de Controle

Para construir uma interface comum entre os dados do movimento capturado e a interface baseada em visão, foi derivado um pequeno conjunto de parâmetros a partir das características faciais rastreadas como sinal de um controle robusto e discriminativo para a animação facial. No total o sistema automaticamente extraiu 15 parâmetros de controle que descrevem a expressão facial de um "ator" observado:

- **Boca** (*06 parâmetros*): o sistema extraiu seis quantidades escalares descrevendo o movimento da boca tendo como base as quatro características rastreadas ao redor da boca: canto esquerdo, canto direito, lábio superior e lábio inferior. Mais precisamente, esses seis parâmetros de controle incluem parâmetros que medem a distância entre os lábios superiores e inferior (1), a distância entre os cantos direito e esquerdo da boca (1), o centro da boca (2), o ângulo do segmento de linha conectando o lábio superior e o lábio inferior com relação à linha vertical (1) e o ângulo do segmento de linha conectando os cantos direito e esquerdo com relação à linha horizontal.
- **Nariz** (*02 parâmetros*): baseado nas três características rastreadas no nariz, o sistema computou dois parâmetros de controle descrevendo o movimento do nariz: a distância entre os cantos direito e esquerdo do nariz (1) e a distância entre o ponto topo do nariz e o segmento de linha conectando os cantos esquerdo e direito (1).
- **Olhos** (*02 parâmetros*): os parâmetros de controle que descrevem os movimentos dos olhos são a distância entre as pálpebras superior e inferior de cada olho (2).
- **Sobrancelha** (*05 parâmetros*): os parâmetros de controle para as ações da sobrancelha consiste do ângulo de cada sobrancelha em relação à linha horizontal (2), a distância entre cada sobrancelha e o olho (2) e a distância entre a sobrancelha direita e esquerda (1).

Esses 15 parâmetros são usados para controlar a deformação de uma expressão em um modelo de personagem e será utilizada a notação $\tilde{Z}_i \equiv \{\tilde{z}_{a,i} \mid a = 1,...,15\}$ para denotar o controle do sinal no tempo *i*, onde *a* é o índice para um parâmetro de controle individual.

8.3. PRÉ-PROCESSAMENTO DOS DADOS DA CAPTURA DE MOVIMENTO

Foi utilizado um *scanner* a laser *Minolta Vivid 700* para construir o modelo de superfície da captura de movimento de uma pessoa [Huber2001]. Foi definido um sistema de captura de movimento *Vicon* para gravar o movimento facial anexando 76 marcas reflectivas na face da pessoa que irá ter seu movimento capturado, como ilustra a Figura 8-4.

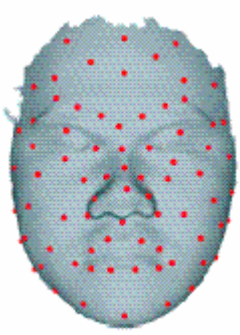


Figura 8-4: Modelo de superfície da cabeça escaneada da captura de movimento de uma pessoa alinhado com 76 marcas de captura de movimento.

Durante as sessões de captura, a pessoa tinha liberdade de mover sua cabeça livremente porque o movimento da cabeça está envolvido na maioria das expressões de forma natural. Como resultado, o movimento da cabeça e o movimento das expressões faciais estão casados nos dados do movimento capturado e precisam ser separados.

8.3.1. Separando Posição da Cabeça e Expressão

Assumindo que a expressão facial deforma com L modos independentes de variação, seu formato pode ser representado como uma combinação linear de um conjunto de bases de deformação $S_1, S_2, ..., S_L$. Cada base de deformação S_i é uma matriz $3 \times P$ descrevendo o modo de deformação de P pontos. Os dados do movimento capturado facial gravado X_f combina os efeitos da posição da cabeça *3D* e a deformação da expressão local: $X_f = R_f \cdot \left(\sum_{i=1}^{L} c_{fi} \cdot S_i \right) + T_f$ onde R_f é uma matriz 3×3 de rotação da cabeça e T_f é uma matriz 3×1 de translação da cabeça no quadro f. c_{fi} é o peso correspondente a i-ésima base de deformação S_i. O objetivo é separar a pose da cabeça

R_f e T_f do dado capturado X_f de tal forma que o dado do movimento capturado gravado tenha apenas deformações de expressão.

T_f foi eliminado de X_f através da subtração da média de todos os pontos *3D*. O dado da captura de movimento resultante foi representado em notação de matriz:

$$\underbrace{\begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_F \end{pmatrix}}_{M} = \underbrace{\begin{pmatrix} c_{11}R_1 & \dots & c_{1L}R_1 \\ & . & \\ \dots & . & \dots \\ & . & \\ c_{F1}R_F & \dots & c_{FL}R_F \end{pmatrix}}_{Q} \cdot \underbrace{\begin{pmatrix} S_1 \\ . \\ . \\ . \\ S_L \end{pmatrix}}_{B}$$

onde F é o número de quadros do dado do movimento capturado, M é uma matriz $3F \times P$ armazenando as coordenadas 3D de todas as posições de marcas do movimento capturado, Q é uma matriz $3F \times 3L$ de rotação escalada armazenando as orientações da cabeça e o peso de cada base de deformação em cada quadro e B é uma matriz $3L \times P$ contendo todas as bases de deformação [Chai2003].

Depois de separar as poses da cabeça do movimento de dados capturados, cada quadro dos dados do movimento capturado é projetado em uma visão frontal-paralela e são extraídos os parâmetros de controle da expressão para cada quadro do movimento capturado. Sejam $X_i \equiv \{x_{b,i} \mid b = 1,...,76\}$ as posições 3D das marcas do movimento capturado no quadro *i* e seja $Z_i \equiv \{z_{a,i} \mid a = 1,...,15\}$ os parâmetros de controle derivados a partir do quadro *i*. Aqui $x_{b,i}$ é a coordenada 3D da *b*-ésima marca de captura de movimento correspondente ao quadro *i*. Dessa forma, cada quadro de captura de movimento X_i é automaticamente associado com os parâmetros de controle da animação Z_i.

8.4. CONTROLE DA EXPRESSÃO E DA ANIMAÇÃO

Dados os parâmetros de controle derivados a partir da interface baseada em visão, controlar o movimento da cabeça de um personagem virtual é direto. O sistema mapeia diretamente a orientação do usuário no personagem virtual. Os parâmetros de posição derivados a partir do vídeo precisam ser escalados apropriadamente antes deles serem utilizados para controlar a posição do avatar. Essa escala é computada como uma taxa da largura da boca entre o usuário e o avatar.

Controlar a deformação da expressão necessita integrar a informação dos parâmetros de controle da expressão e dados da captura de movimento. Nesta seção, o artigo apresenta uma nova abordagem dirigida por dados para síntese de movimento que translada o ruído e os sinais de controle da expressão com baixa resolução para dados de movimento com alta resolução usando a informação contida na base de dados de captura de movimento.

8.4.1. Normalização dos Parâmetros de Controle

Os parâmetros de controle da expressão a partir dos dados do rastreamento (*tracking*) são inconsistentes com os parâmetros de controle dos dados da captura de movimento porque o usuário e a pessoa da captura de movimento possuem proporções faciais e geometria facial diferentes. O artigo fez uso de um passo de normalização que automaticamente escala a medida dos parâmetros de controle de acordo com a expressão neutra para, aproximadamente, remover essas diferenças.

Através dessa escala dos parâmetros de controle, é possível afirmar que os parâmetros de controle extraídos do usuário têm, aproximadamente, a mesma magnitude daqueles extraídos dos dados de captura de movimento quando ambos estão na mesma expressão facial.

8.4.2. Filtragem Dirigida pelos Dados

Sinais de controle em uma interface baseada em visão normalmente possuem ruído. O trabalho do artigo dividiu esses sinais em segmentos com um tamanho temporal pequeno e fixo, rotulado de *W*, em tempo de execução, e depois utilizou conhecimento embutido na base de dados de captura de movimento para seqüencialmente filtrar os sinais de controle segmento por segmento. O primeiro modelo dos sinais de controle da expressão é um modelo linear local aprendido em tempo de execução. Quando um novo segmento chega, a base de dados de captura de movimento procura por exemplos que são mais relevantes para esse segmento. Esses exemplos são então utilizados para construir um modelo local dinamicamente linear que captura o comportamento dinâmico dos sinais de controle sobre a seqüência de tamanho fixo.

Um conjunto de quadros vizinhos com a mesma janela temporal *W* é coletado para cada quadro na base de dados de captura de movimento e eles são tratados como um dado pontual. Conceitualmente, todos os segmentos de movimento dos sinais de controle facial formam um *mainfold* não-linear embutido em um espaço de configuração de alta dimensão.

A idéia chave da técnica de filtragem desenvolvida no artigo diz que é possível utilizar um subespaço linear de baixa dimensão para aproximar a região local do *mainfold* não-linear de alta dimensão. Para cada amostra de ruído foi aplicada análise de componentes principais (PCA – *Principles Components Analysis*) para aprender um subespaço linear usando os dados pontuais que caem dentro da região local e depois eles são reconstruídos usando subespaços lineares de baixa dimensão.

Os tamanhos dos segmentos de movimento irão determinar o tempo de resposta do sistema, ou o atraso de ação-reação entre o usuário e o avatar. A partir dos experimentos desenvolvidos, constatou-se que vinte é um número razoável para a quantidade de quadros para cada segmento. O atraso da ação entre o usuário e o avatar foi 0,33s porque a taxa de quadro da câmera de vídeo utilizada foi de 60fps.

Seja $\widetilde{\phi}_t \equiv \left[\widetilde{Z}_1,...,\widetilde{Z}_W\right]$ um fragmento dos parâmetros de controle de entrada. O passo do filtro é:

- Encontrar as *K* fatias mais próximas da base de dados da captura de movimento;
- Computar os componentes principais das *K* fatias mais próximas. Foram guardados os *M* maiores autovetores $U_1,...,U_M$ como bases de filtros, e *M*

foi automaticamente determinado através da retenção de 99% da variação dos dados originais; e

- Projetar $\widetilde{\phi}_t$ em um espaço linear local expandido por $U_1,...,U_M$ e reconstruir o sinal de controle $\widetilde{\phi}_t \equiv \left[\widetilde{Z}_1,...,\widetilde{Z}_b\right]$ usando os coeficientes da projeção.

Os componentes principais, que podem ser interpretados como origens de maior variação dinâmica na região local dos sinais de controle da entrada, naturalmente capturam o comportamento dinâmico dos parâmetros de controle da animação na região local do sinal de controle da entrada $\widetilde{\phi}_t$. A escolha específica do *K* depende das propriedades da base de dados da captura de movimento fornecida e dos dados de controle extraídos a partir do vídeo. Nos experimentos realizados, verificou-se que um *K* entre 50 e 150 gerava bons resultados de filtragem.

8.4.3. Síntese de Expressão Dirigida pelos Dados

Supondo *N* quadros de dados de captura de movimento $X_1,...,X_N$ e associando esse quadros com parâmetros de controle $Z_1,...,Z_N$, o problema de controle de expressão pode ser estabelecido da seguinte forma: dado um segmento de sinal de controle $\overline{\phi}_t = \left[\overline{Z}_{t_1},...,\overline{Z}_{t_W}\right]$ deseja-se sintetizar o exemplo de captura de movimento correspondente $\overline{M}_t = \left[\overline{X}_{t_1},...,\overline{X}_{t_W}\right]$.

A solução mais simples para esse problema pode ser a interpolação dos *K*-ésimos vizinhos mais próximos. Para cada quadro $\overline{Z}_i$ tal que $i = t_1,...,t_W$ o sistema pode escolher os *K* exemplos mais próximos dos pontos de parâmetros de controle, denotados como $Z_{i_1},...,Z_{i_K}$, nos dados de captura de movimento e atribuir a cada um deles um peso ω_i baseado na distância deles à fila $\overline{Z}_i$. Esses pesos podem ser depois aplicados para sintetizar os dados do movimento *3D* através da interpolação dos dados correspondentes de captura de movimento $X_{i_1},...,X_{i_K}$. A desvantagem dessa abordagem é que o movimento gerado pode não ser suave devido ao fato que o mapeamento do espaço do parâmetro de controle para o espaço de configuração do movimento não é um para um.

Uma vez que isso pode acontecer, foi desenvolvida uma técnica de síntese de movimento baseado em segmento para, justamente, remover a ambigüidade do mapeamento a partir do espaço de parâmetros de controle para o espaço de configuração do movimento. Dada a fila de segmento $\overline{\phi}_t$, é computado os pesos da interpolação baseado na sua distância aos *K*-ésimos segmentos mais próximos da base de dados de captura de movimento. A distância é medida como uma distância Euclidiana em um subespaço linear local. É possível sintetizar o segmento dos dados de movimento através de uma combinação linear dos dados de movimento nos *K* segmentos mais próximos. Sintetizando o movimento dessa forma torna-se possível gerenciar a ambigüidade no mapeamento através da integração dos parâmetros de controle da expressão mais adiante e mais posterior em todo o segmento.

8.4.4. Estrutura dos Dados

Um grande número de *K* filas de vizinhos mais próximo precisa ser conduzido sobre um mesmo conjunto de dados *S*. O custo computacional dessa ação pode ser reduzido se houver um pré-processamento de *S* para criar uma estrutura de dados que permita fazer uma busca mais rápida dos pontos mais próximos.

O artigo apresenta uma estrutura de dados e uma técnica de busca eficiente para os *K* vizinhos mais próximos que faz uso da vantagem de coerência temporal dos dados da fila, ou seja, exemplos vizinhos na fila estão dentro de uma distância menor no espaço de alta-dimensão. A busca pelo *K* ponto mais próximo pode ser descrita pelos seguintes passos:

- Construir um grafo de vizinhança *G*: são coletados o conjunto de quadros vizinhos com uma janela temporal de *W* para cada quadro na base de dados de captura de movimento e eles são tratados como um dado pontual. Com isso é computada a distância Euclidiana padrão $d_x(i, j)$ para cada par *i* e *j* dos dados pontuais na base de dados. Um grafo *G* é definido sobre todos os dados pontuais através da concatenação dos pontos *i* e *j* se eles forem próximos quanto o determinado por um limiar ε, ou se *i* for um dos *K* vizinhos mais próximos de *j*. Os tamanhos da arestas de *G* são definidos por $d_x(i, j)$. O grafo gerado é usado como uma estrutura de dados para filas eficientes dos pontos mais próximos.
- Busca do *K* vizinho mais próximo: quando um novo movimento chega na fila, primeiro encontra-se o exemplo *E* mais próximo entre os *K* pontos mais próximos da última busca no *buffer* [Chai2003].

Uma única busca em uma base de dados de tamanho |*S*| pode ser aproximadamente alcançada em tempo *O(K)*, que é independente do tamanho da base de dados |*S*| e é muito mais rápido do que uma busca exaustiva com complexidade de tempo linear *O(|S|)*.

8.5. RE-ALVO DAS EXPRESSÕES

Nesta seção será apresentado um método eficiente de re-alvo (*retargetting*) de expressão cuja computação em tempo de execução é constante independente da complexidade do modelo do personagem. A idéia básica desse método de re-alvo das expressões é pré-computar todas as bases de deformações do modelo alvo de tal forma que a operação em tempo de execução envolve apenas combinar essas bases de deformação apropriadamente.

A entrada para esse processo é o modelo da superfície origem escaneada, o modelo da superfície destino (alvo) e as bases de deformação da base de dados de captura de movimento. É necessário que ambos os modelo (origem e destino) estejam na expressão natural (neutra).

O sistema primeiro constrói a superfície correspondente entre os dois modelos e então adapta as bases de deformação do modelo origem para o modelo destino a partir do comportamento de deformação derivado de cada superfície correspondente. Em tempo de execução, o sistema opera no movimento sintetizado para gerar o peso para cada base de deformação. A animação final é criada através da combinação das bases de deformação

usando os pesos. Em particular, o processo de re-alvo das expressões consiste de quatro estágios: *motion vector interpolation* (interpolação do vetor movimento), *dense surface correspondences* (correspondências da superfície densa), *motion vector transfer* (transferência do vetor de movimento) e *target motion synthesis* (síntese do movimento alvo).

8.5.1. Interpolação do Vetor de Movimento

Dado o vetor de deformação dos pontos-chave da superfície origem para cada modo de deformação, esse passo tem como objetivo deformar os vértices restantes da superfície origem através da interpolação linear do movimento dos pontos-chave usando coordenadas do baricentro. Primeiro o sistema gera a malha do modelo baseado nas posições *3D* das marcas de captura de movimento no modelo origem. Para cada vértice, o sistema determina a face na qual o vértice está localizado e as coordenadas baricêntricas para a interpolação. Então o vetor de deformação dos vértices restantes é interpolado de acordo.

8.5.2. Correspondências da Superfície Densa

Começando com um pequeno conjunto de correspondências manualmente estabelecidas entre duas superfícies, uma superfície densa correspondente é computada através do *morphing* do volume com uma Função de Base Radial seguida de uma projeção cilíndrica, como ilustra a Figura 8-5. Essa correspondência fornece um mapeamento homeomórfico (um para um e em um) entre as duas superfícies.

Depois foi utilizado Funções de Base Radial, mais uma vez, para aprender o mapeamento homeomórfico contínuo da função $f(x_S) = (f_1(x_S), f_2(x_S), f_3(x_S))$ da expressão neutra da superfície origem $x_S = (x_S, y_S, z_S)$ para a expressão neutra da superfície destino $x_t = (x_t, y_t, z_t)$, tal que $x_t = f(x_S)$.

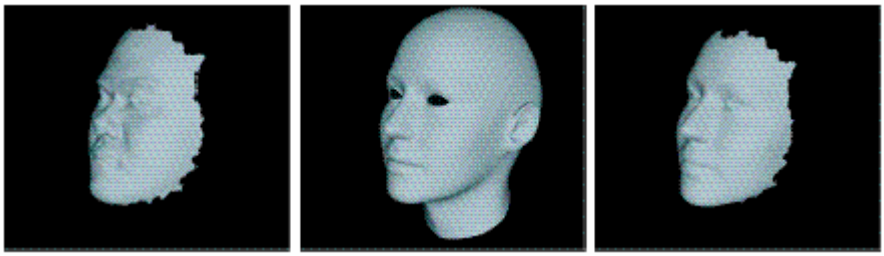


Figura 8-5: Correspondência das superfícies densas: mais à esquerda o modelo da superfície origem escaneada, ao meio o modelo da superfície animada e mais à direita o modelo *morphed* da superfície origem com a superfície destino usando a superfície de correspondência.

8.5.3. Transferência do Vetor de Movimento

A deformação na superfície origem não pode ser simplesmente transferida para o modelo que haja um ajuste da direção e da escala de cada vetor movimento. Isto é necessário porque as proporções faciais e a geometria facial variam entre os modelos. Para cada ponto x_S na superfície de entrada é possível computar o ponto correspondente x_t na superfície destino (alvo) usando o mapeamento de função $f(x_S)$.

Dada uma pequena deformação $\delta x_S = (\delta x_S, \delta y_S, \delta z_S)$ para cada ponto na origem x_S, a deformação δx_S dos pontos correspondentes no destino x_t é computado através de matrizes Jacobianas $Jf(x_S) = (Jf_1(x_S), Jf_2(x_S), Jf_3(x_S))^T : \delta x_t = Jf(x_S) \cdot \delta x_S$.

Foi utilizada a função de aprendizado RBF $f(x_S)$ para computar numericamente a matriz Jacobiana em x_S:

$$Jf_i(x_S) = \begin{pmatrix} \dfrac{f_i(x_S + \delta x_S, y_S, z_S) - f_i(x_S, y_S, z_S)}{\delta x_S} \\ \dfrac{f_i(x_S, y_S + \delta y_S, z_S) - f_i(x_S, y_S, z_S)}{\delta y_S} \\ \dfrac{f_i(x_S, y_S, z_S + \delta z_S) - f_i(x_S, y_S, z_S)}{\delta z_S} \end{pmatrix}, i = 1,2,3$$

Geometricamente, a matriz Jacobiana ajusta a direção e a magnitude do vetor de movimento origem de acordo com a correspondência de superfície local entre os dois modelos. Devido ao fato que a deformação do movimento origem é representado por uma combinação linear de um conjunto de pequenas bases de deformação, a deformação δx_t pode ser computada como: $\delta x_t = Jf(x_S) \cdot \sum_{i=1}^{L} \lambda_i \delta x_{s,i} = \sum_{i=1}^{L} \lambda_i (Jf(x_S) \cdot \delta x_{S,i})$; onde $\delta x_{S,i}$ representa uma pequena deformação correspondendo a *i*-ésima base de deformação e λ_i é o peso correspondente.

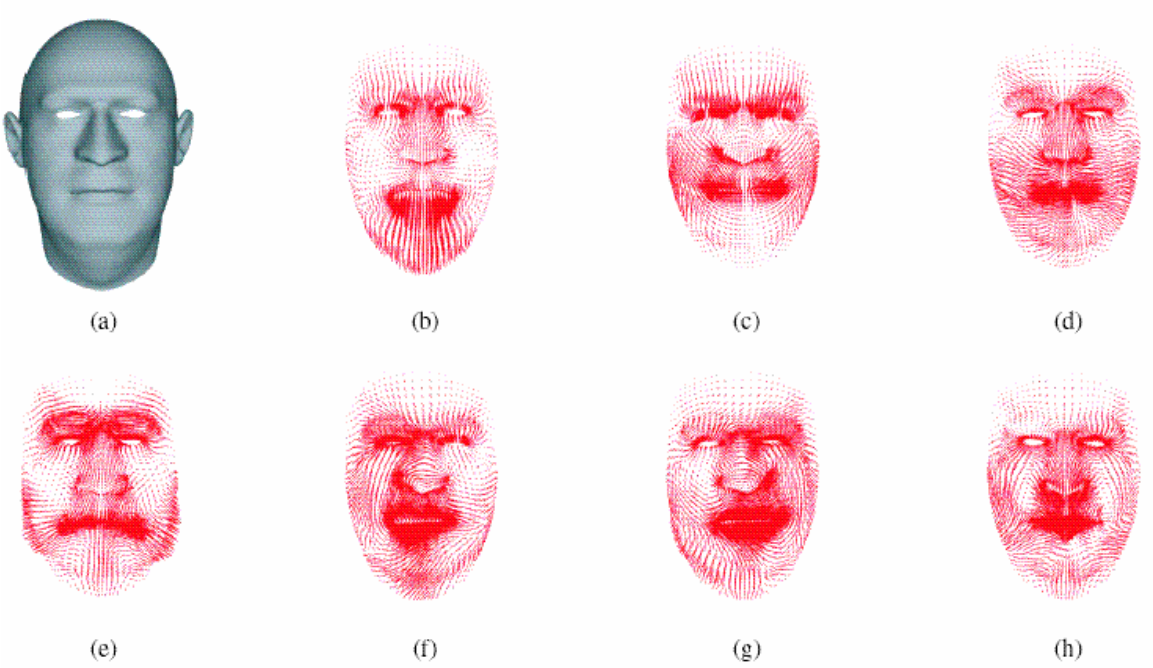


Figura 8-6: As sete bases de deformação para o modelo de superfície destino. Em (a) a máscara cinza do modelo da superfície alvo, e de (b) a (h) as agulhas mostram a escala e a direção do vetor de transformação 3D para cada vértice.

8.5.4. Síntese do Movimento Alvo

Depois dos dados de movimento terem sido sintetizados a partir da base de dados da captura de movimento via a interface baseada em visão, o sistema projeta-os no espaço da base de deformação do modelo origem $S_1,...,S_L$ para computar a combinação de pesos $\lambda_1,...,\lambda_L$. A deformação da superfície destino (alvo) δx_t é gerada pela combinação de todas as bases de deformação da superfície destino $\delta x_{t,i}$ usando os pesos combinados λ_i de acordo com a equação:

$$\delta x_t = Jf(x_S)\cdot\sum_{i=1}^{L}\lambda_i\delta x_{s,i} = \sum_{i=1}^{L}\lambda_i\left(Jf(x_S)\cdot\delta x_{S,i}\right).$$

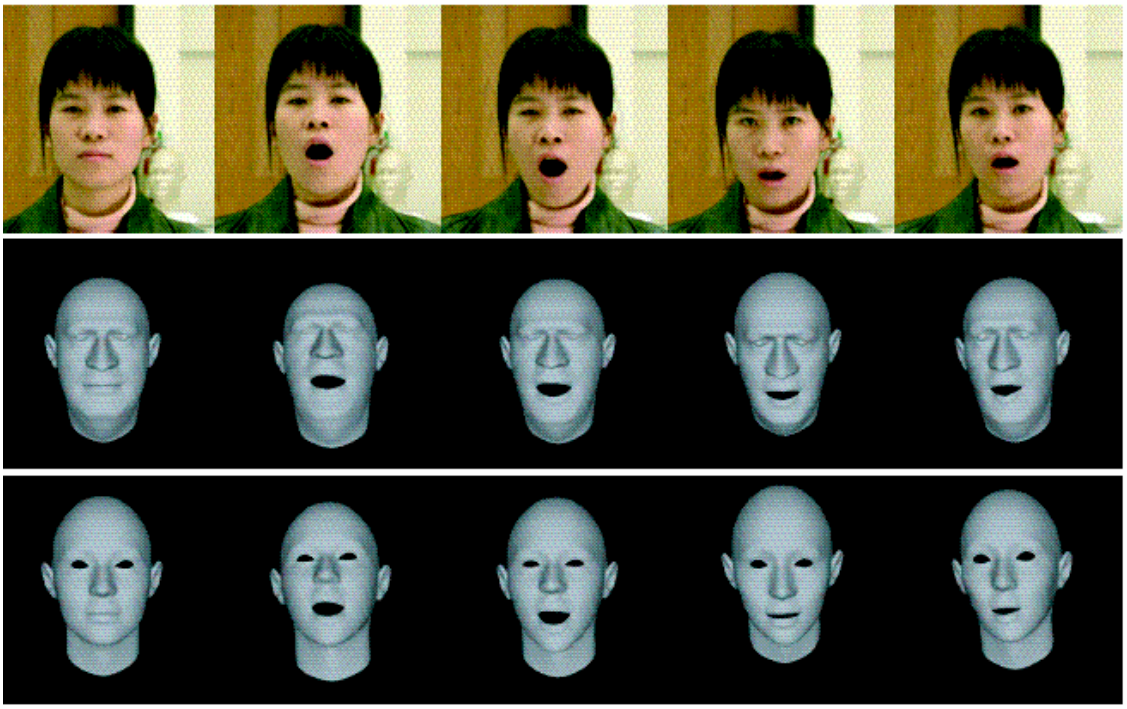
A síntese do movimento alvo é feita *online* (em tempo real) e os outros três passos, previamente apresentados (seções 8.5.1, 8.5.2 e 8.5.3), são completamente *offline*. O custo computacional do tempo de execução do re-alvo da expressão depende apenas do número de bases de deformação para a base de dados de captura de movimento *L*, ao invés de depender do número de vértices no modelo animado. Na implementação desenvolvida no artigo [Chai2003], o custo do re-alvo da expressão é bem menor do que o custo de renderização.

8.6. Resultados e Conclusões do Artigo

Todos os dados de movimento nos experimentos desenvolvidos no artigo foram de uma pessoa e coletados a uma taxa de 120 quadros/segundo. No total, a base de dados de

captura de movimento contém cerca de 70000 quadros (cerca de 10minutos). Foram capturados vários conjuntos de ações faciais, incluindo as seis expressões faciais: *anger* (com raiva), *fear* (com medo), *surprise* (surpresa), *sadness* (tristeza), *joy* (alegria) e *disgust* (aversão), além de outras ações comuns como comendo, bocejando e ressonando. Foi também gravada uma quantidade pequena de dados de movimento relacionados à fala (cerca de 6000 quadros), mas a quantidade de dados de fala não foi suficiente para cobrir todas as variações dos movimentos faciais relacionados à fala.

O sistema de animação facial desenvolvido no artigo foi testado por diferentes usuários. Foi utilizada uma única câmera de vídeo para gravar a expressão facial de um ser humano. Esse usuário sendo gravado não tinha conhecimento da base de dados previamente montada, como também não tinha restrições quanto aos seus movimentos. O usuário era apenas instruído de começar a partir da expressão natural (neutra) e estando em uma posição (visão) paralela-frontal à câmera de vídeo. A Figura 8-7e a Figura 8-8 ilustram vários exemplos de amostras de quadros a partir de uma única câmera de vídeo de dois usuários e a animação facial 3D das expressões de dois personagens virtuais diferentes.



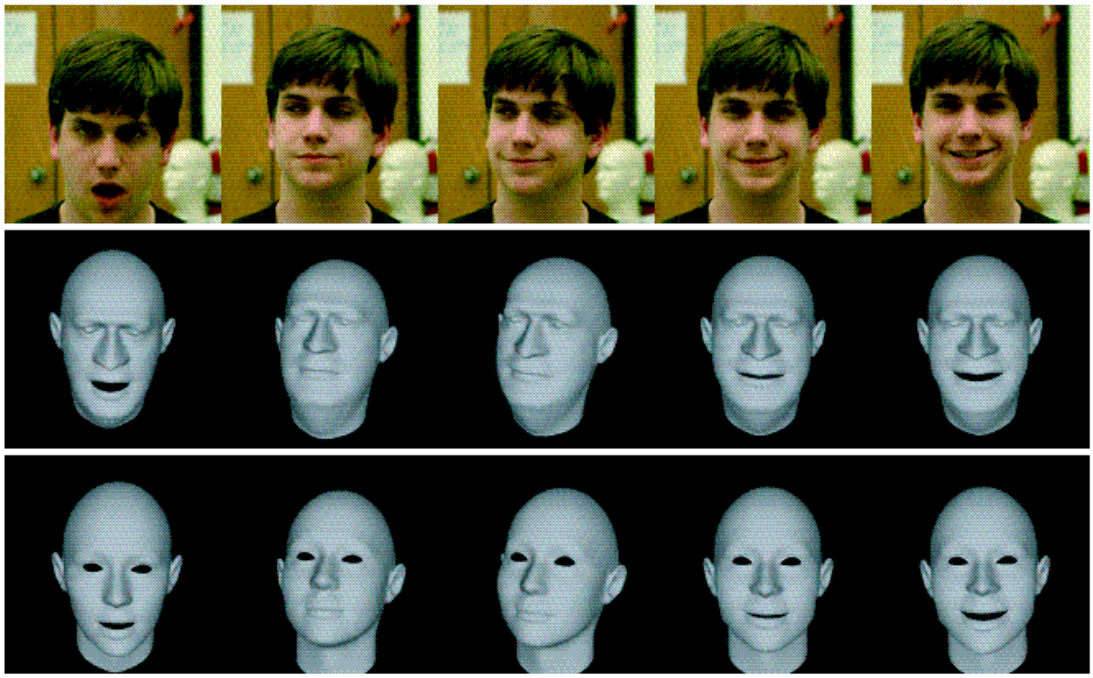


Figura 8-7: Resultado de dois usuários controlando e animando expressões faciais 3D de dois modelos diferentes de superfície destino (alvo).

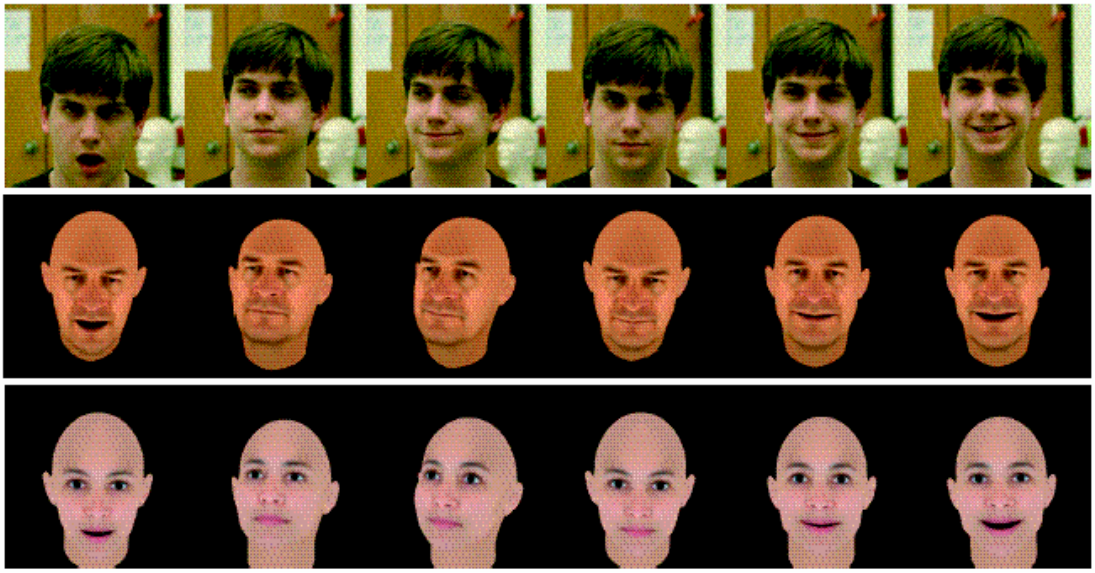


Figura 8-8: Resultado de dois usuários controlando e animando expressões faciais 3D de dois modelos diferentes de avatar com mapeamento de textura.

O artigo, de modo geral, demonstrou como um pré-processamento de captura de movimento em conjunto com um sistema de rastreamento facial em tempo real pode ser utilizado para criar uma animação facial baseada em performance onde um usuário (ser humano) efetivamente controla as expressões e as ações faciais de um personagem virtual 3D. Em particular, foi desenvolvido um sistema de animação fim-a-fim para rastreamento de movimento facial em tempo real em um vídeo, pré-processando uma base de dados de captura de movimento, controlando e animando as ações faciais usando a base de dados de captura de movimento e extraindo parâmetros de controle da animação e fazendo o re-alvo das expressões sintetizadas em um novo modelo alvo.

Uma limitação apontada no artigo desse sistema desenvolvido é que algumas vezes ele perde detalhes do movimento labial. Esse problema acontece porque a base de dados de captura de movimento não incluiu amostras suficientes relacionadas à fala. Alternativamente, a quantidade de dados rastreados na interface baseada em visão não foi suficiente para capturar os movimentos sutis dos lábios.

Atualmente, o sistema desenvolvido não considera a fala como entrada. A combinação de uma interface baseada em fala com uma interface baseada em visão poderia melhorar bastante a qualidade da animação final e melhorar o controle dos movimentos faciais.

Um outro ponto interessante de estudo é a exploração de outras interfaces intuitivas para animação facial dirigida por dados. Por exemplo, um sistema de interpolação *key-framing* poderia usar as informações contidas na base de dados de captura de movimento para adicionar detalhes do movimento vivo aos graus de liberdade que não são especificados por um animador.

8.7. CONCLUSÕES PESSOAIS

i. Quem são os autores?

Jin-xiang Chai: Aluna do quarto ano de doutorado do Instituto de Robótica da Escola de Ciência da Computação na *Carnegie Mellon University*. Vem desenvolvendo pesquisa na área de controle interativo de avatar, mais especificamente, ela vem desenvolvendo algoritmos de animação que gerar ações vivas para personagens 3D interativos. Sua orientadora é a Profa. Jessica Hodgins. Basicamente, este artigo é fruto de seu trabalho de doutorado.

Jing Xiao: Aluno de doutorado do Instituto de Robótica da Escola de Ciência da Computação na *Carnegie Mellon University*. Ele vem trabalhando com os professores Takeo Kanade e Jeffrey Cohn nas áreas de Visão Computacional e Reconhecimento de Padrões.

Jessica Hodgins: Professora de Ciência da Computação e Robótica da *Carnegie Mellon University*. Basicamente sua pesquisa vem sendo nas áreas de robótica humana, animação humana, animação passiva e comportamento de grupo.

ii. O que o artigo resolve?

O artigo apresentou o desenvolvimento de um sistema em que um usuário humano qualquer faz uma ação ou expressão facial e um avatar "imita" essa expressão ou ação em tempo real.

iii. Qual a abordagem utilizada?

O sistema desenvolvido fez uso de várias abordagens (sub-etapas) para o seu desenvolvimento. Basicamente o sistema utilizou captura de movimento para montar uma base de dados de onde o avatar aprende as ações que ele pode fazer. O sistema também usou a abordagem de *tracking* para poder capturar pontos característicos da face do usuário e assim mapear a ação/expressão no avatar. Uma outra abordagem também foi o uso de uma interface baseada em visão onde a expressão/ação do usuário era capturada através de uma única câmera de vídeo.

iv. Qual a classificação do artigo?

O artigo caracteriza-se por ser ao mesmo tempo de *tracking* e de captura de expressões faciais/ações.

v. Quais foram as ferramentas utilizadas na implementação?

O artigo não comenta a linguagem de implementação utilizada no desenvolvimento do sistema, nem muito mesmo o processador utilizado nos testes e no desenvolvimento. O único comentário do artigo em relação a essas características é do *scanner* a laser Minolta Vivid 700 utilizado para "escanear" a superfície do modelo de face da pessoa da captura de movimento para construção da base de dados; como também da utilização de uma câmera de vídeo como fonte da interface baseada em visão.

vi. Quais os problemas em aberto interessantes que o artigo apresenta ("ataca")?

O artigo é bastante completo. A idéia parece ser bastante interessante além de fazer uso do trabalho em conjunto de diferentes abordagens, cada uma contribuindo com uma de suas vantagens. Particularmente, a idéia de interface baseada em visão é bastante interessante, tanto no que foi aplicado pelo artigo, como também para se ter uma "resposta" do avatar não "imitando a ação do usuário" e sim "respondendo/complementando a ação do usuário", pensando também na possibilidade de se ter uma base de dados de captura de movimento como fonte de aprendizado para o avatar. A idéia de ter um mapeamento de textura também é bastante interessante, onde é possível se pensar em uma textura foto-realista.

vii. O que não foi feito neste artigo que é interessante? (Quais os problemas em aberto interessantes que o artigo desperta e não ataca?)

Apesar do artigo mencionar ações de fala não foi possível refletir de forma consistente e convincente essa ação no avatar. O artigo menciona como um dos trabalhos futuros essa investigação da fala como interface de entrada. A partir desse ponto seria interessante verificar como seria feito o mapeamento dos visemas tendo a fala e as expressões em conjunto.

9. Conclusões

Como comentado no capítulo introdutório deste documento, esta monografia tem como objetivo apresentar um resumo de alguns dos principais artigos, recentemente publicados, da área de animação facial. Os artigos apresentados foram publicados em conferências importantes da área e possuem em comum a busca por desenvolver uma animação facial o mais realista e convincente possível.

A Figura 9-1 ilustra a face trabalhada no primeiro artigo apresentado [Bui2003]. O artigo apresentou o desenvolvimento de uma malha facial tridimensional e fez uso da abordagem baseada em músculos para a animação da face. Apenas em alguns pontos mais críticos de modelagem e conseqüente animação, como os olhos e a boca, foi utilizada uma abordagem de pseudo-músculo.

A partir da expressão de neutralidade e de aplicação de contrações musculares é possível, em tempo real, formar novas expressões faciais. O artigo afirma que suas expressões faciais são realistas e sua animação (translação de uma expressão facial para outra) ocorre de forma rápida, até mesmo se executada em um computador pessoal.

Um ponto falho neste artigo e que poderia ser mais bem explorado é a associação do modelo 3D desenvolvido a um sistema *talking head*. Como o próprio artigo menciona, ele não se preocupa com a formação dos visemas e, conseqüentemente, não aborda as limitações e problemas que ocorrem quando do casamento (sobreposição) dos visemas com as expressões faciais que movimentam os lábios de seu estado de neutralidade.

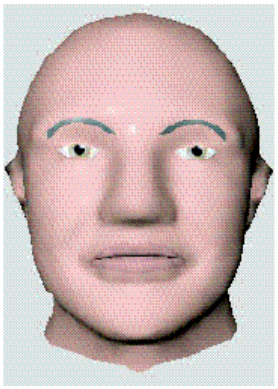


Figura 9-1: Face desenvolvida no primeiro artigo apresentado

A Figura 9-2 ilustra os pontos característicos e a divisão da face em regiões, sendo essas algumas das contribuições do segundo artigo apresentado [Zhang2003]. O artigo fez uso da abordagem baseada em performance para animar o modelo de face do trabalho.

O artigo apresenta um sistema que, a partir de imagens dadas como entrada, exibe a mesma expressão facial da imagem em um novo personagem sintetizado. O artigo também descreve um editor de expressões onde o usuário pode, interativamente, movimentar pontos característicos da face e ter, como resposta em tempo real, as novas expressões faciais que o próprio usuário está definindo.

Assim como no primeiro artigo apresentado [Bui2003], o trabalho preocupa-se com a produção de uma expressão facial realista. No entanto, para conseguir isto ele precisa abstrair da emoção facial casada com a fala. Os autores apontam a "emoção facial dirigida pela fala" como um dos possíveis trabalhos futuros.

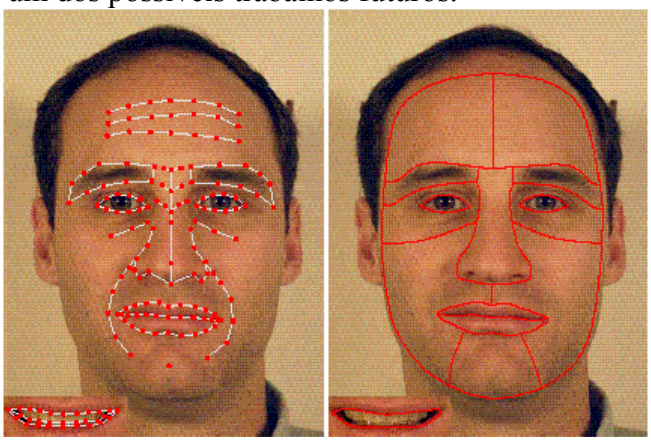


Figura 9-2: Pontos característicos e divisão da face em regiões, trabalhada no artigo II.

A Figura 9-3 ilustra o cenário desenvolvido para os experimentos do terceiro artigo apresentado neste documento [Cunningham2003]. A idéia do artigo foi desenvolver experimentos para avaliar quão verdadeiras e convincentes eram as 06 expressões universais [Ekman1971]. A infra-estrutura montada para o experimento parece ser bastante eficiente e como o próprio artigo aponta, existem várias outras experiências a serem desenvolvidas a partir dessa primeira.

O artigo comenta que a partir desses experimentos seria possível verificar quais estruturas faciais são mais importantes para a identificação das expressões faciais. Isso foi comentado, mas não apareceu nenhuma contribuição formal a esse respeito, que seria um ponto bastante importante a ser explorado.

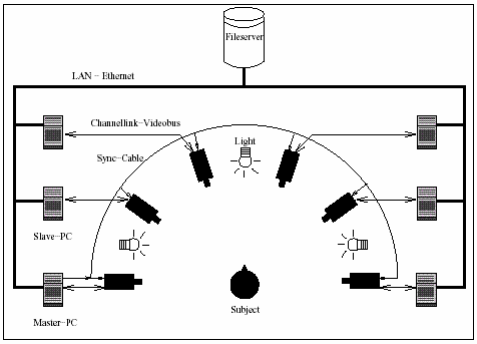


Figura 9-3: Cenário da experiência desenvolvida no artigo III.

A Figura 9-4 ilustra os componentes independentes extraídos a partir da aplicação de ICA nos sinais de fala e expressão facial fornecidos como entrada para o sistema desenvolvido no quarto artigo apresentado [Cao2003]. Basicamente, o artigo define uma ferramenta de edição da fala, onde o sinal de fala é separado em duas "componentes"

principais: o conteúdo da fala propriamente dita e a emoção associada com a fala em questão.

Um ponto interessante deste artigo é o uso de análises de componentes independentes (ICA) para a separação do componente de emoção do componente da fala propriamente dita. Abordagens estatísticas vêm sendo bastante utilizadas na área de animação facial e essa utilização tem proporcionado resultados bastante satisfatórios. Por fim, uma crítica ao artigo é que apesar de conseguir fazer a separação dos componentes fala e emoção, a cada momento apenas um deles é aplicado na região da boca (aquele de maior importância).

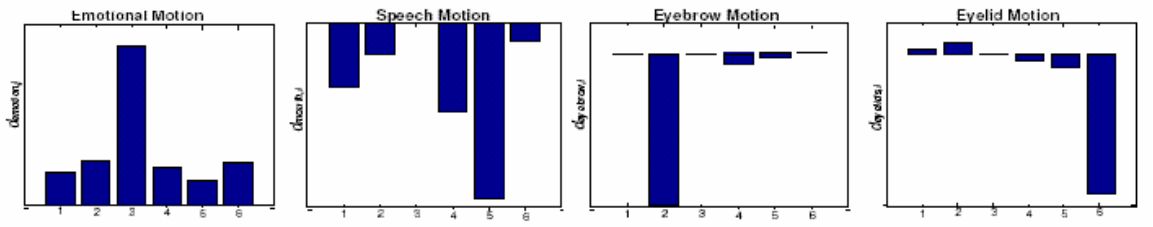


Figura 9-4: Componentes independentes frutos da aplicação de ICA nos resultados do artigo IV.

A Figura 9-5 ilustra uma das principais contribuições do quinto artigo apresentado [Joshi2003]. A partir de *blend shapes* é possível extrair parâmetros de controle para uma animação facial realista. O artigo apresenta uma nova técnica de segmentação da face bastante interessante e que, realmente, parece ser bastante eficiente.

Com a segmentação e os *blend shapes* é possível fazer a animação do personagem tanto através da captura de movimento quanto através da animação de quadros-chave (animação *keyframing*). O artigo trata apenas de expressões faciais, não fazendo alusão ao casamento dessas com a fala, que poderia ser uma abordagem bastante interessante.

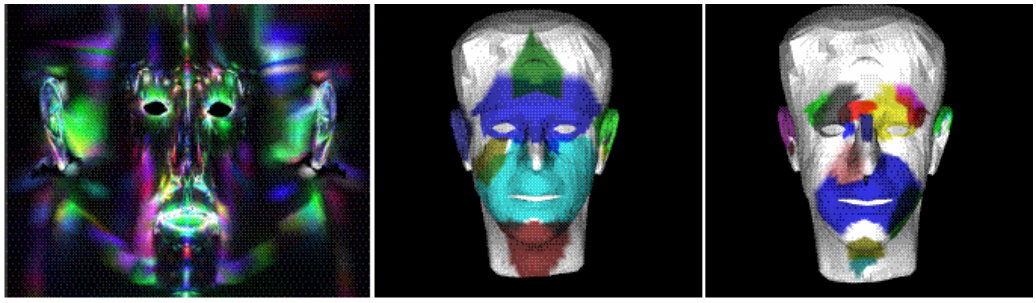


Figura 9-5: *Blend-Shapes* do artigo 5 com segmentações utilizando dois limiares.

A Figura 9-6 ilustra uma visão geral do sistema desenvolvido no sexto artigo apresentado [Pyun2003]. Esse artigo apresenta uma nova técnica de clonagem de expressões faciais de um modelo origem para um modelo destino, preservando as características faciais (geometria e proporções) desse modelo alvo.

Apesar do artigo definir uma métrica de importância para cada vértice da malha facial que permite identificar se a importância em questão é prioritariamente do tipo emocional ou do tipo verbal, não é feito o casamento dessas importâncias caso um determinado trecho de fala tenha emoção. O que o artigo estabelece é que se um certo

vértice tiver uma determinada importância prevalece a expressão facial naquele vértice, caso contrário, é aplicado o visema (contribuição verbal).

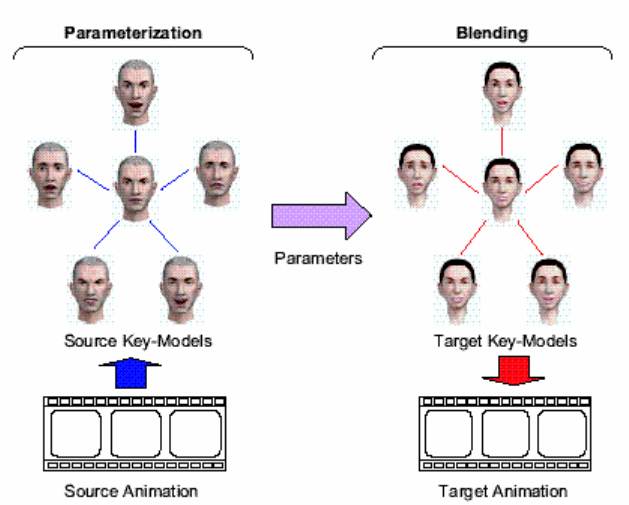


Figura 9-6: Visão geral do sistema de clonagem de expressões faciais desenvolvido no artigo VI.

Por fim, a Figura 9-7 ilustra o resultado do trabalho desenvolvido no sétimo artigo apresentado [Chai2003]. O artigo descreve um sistema em que um usuário humano qualquer faz uma ação ou expressão facial e um avatar "imita" essa expressão ou ação em tempo real. Como nos artigos anteriormente apresentados, apesar de mencionar ações de fala, não foi possível refletir de forma consistente e convincente essa ação no avatar.

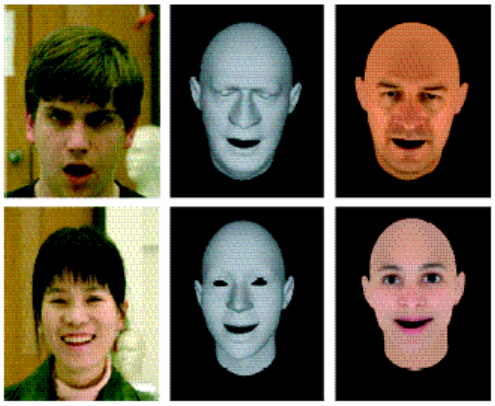


Figura 9-7: Resultado do trabalho desenvolvido no artigo VII.

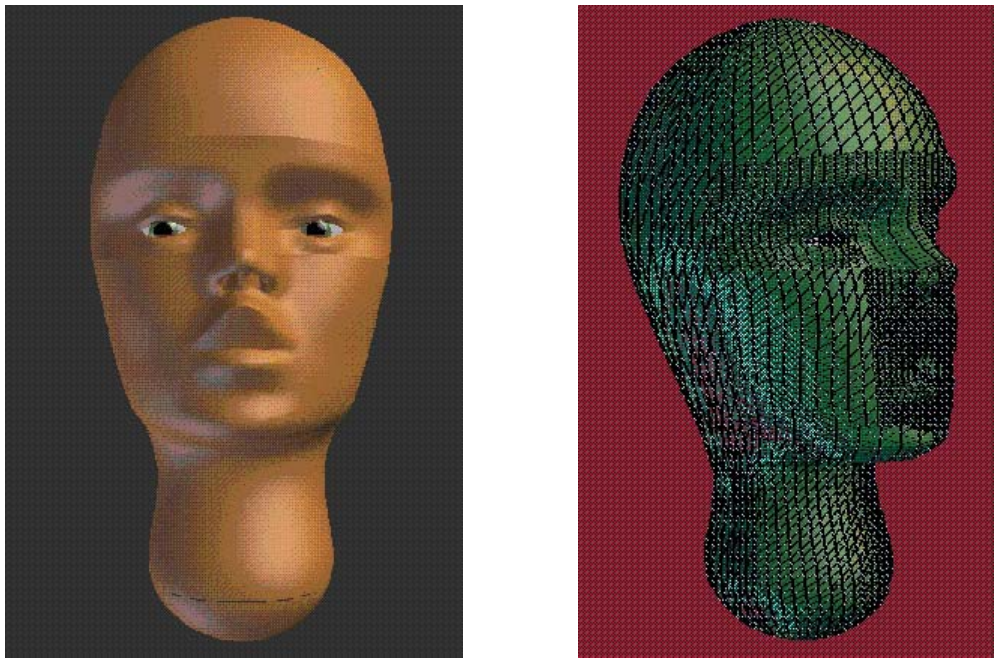
9.1. TRABALHOS FUTUROS

O interessante no desenvolvimento desta monografia foi a possibilidade de extrair um pouco (ou muito) de cada um dos trabalhos apresentados para o trabalho da autora. Como já comentado, a produção de uma animação facial de alta qualidade é um dos problemas mais desafiadores na animação por computador. Centenas de músculos

contribuem para a geração de expressões faciais complexas e para a fala [Parke1996] [Lucena2002].

Particularmente, um problema que ainda parece se encontrar em aberto é o casamento das expressões faciais com a fala em um sistema de *talking head* de animação facial. Como afirmado desde [Parke1996] e ainda relatado em publicações mais recentes como em [Cao2003], nenhum sistema desenvolvido é capaz de simular faces realistas em tempo real e, mais ainda, tendo as expressões faciais e a fala trabalhando em conjunto.

Um ponto de partida interessante para a busca da solução desse desafio poderia ser utilizando o modelo facial 3D desenvolvido pelos professores Luiz Velho[7] e Ken Perlin[8] [Perlin2003], como ilustra a Figura 9-8. A partir dessa malha tridimensional desenvolvida pelo professor Ken Perlin, que constitui de um *applet* na linguagem Java, e já fazendo uso da suavização aplicada pelo Prof. Luiz Velho, poderíamos pensar em animar esse modelo segundo um personagem realista. A suavização desenvolvida é bastante interessante porque ela já leva em consideração as partes da face que possuem uma maior expressividade (região dos olhos e região da boca) e, conseqüentemente, precisam de um modelo poligonal mais refinado (que é um pouco semelhante com o trabalho apresentado em [Bui2003]).



(a) (b)

Figura 9-8: Em (a) Vista frontal da face tridimensional desenvolvida e em (b) vista lateral da malha poligonal da face.

Tendo em mente o desenvolvimento de um personagem foto-realista, poderia ser aplicada uma abordagem de *tracking* para rastrear pontos característicos da malha

[7] Prof. Luiz Velho – IMPA: http://www.impa.br/~lvelho.
[8] Prof. Ken Perlin – MiraLab NYU: http://mrl.nyu.edu/~perlin/ .

poligonal em uma face capturada a partir de uma fotografia como textura, como o apresentado em [Zhang2003], ou fazer uso de um sistema de interface baseado em visão como o apresentado no artigo [Chai2003].

Tomando como ponto de partida a face 3D desenvolvida pelos Professores Ken Perlin e Luiz Velho, e tendo conhecimento das técnicas de animação facial apresentadas em [Parke1996] e resumidas na Seção 1.2, uma questão interessante é definir qual a melhor abordagem a ser utilizada para a animação deste modelo. Naturalmente, um primeiro passo seria fazer uso de uma abordagem baseada em pseudo-músculo, pois para as pequenas animações que já estão disponíveis em [Perlin2003] essa é a abordagem utilizada.

Uma vez tendo a face modelada e animada segundo um personagem realista, um ponto em aberto e interessante é o desenvolvimento de um sistema *talking head* utilizando a mesma. A experiência obtida em [Lucena2002] leva a questionar a investigação de outros sintetizadores além do "*Festival Speech Synthesis"* [Black2004]. O trabalho em conjunto do *Festival* e do MBROLA [Dutoit1998] fez com que o "*Expressive Talking Heads*" [Lucena2002] pudesse ser um sistema bastante flexível, mas na prática o *Festival* demonstrou ser um pouco difícil para a implementação de outros idiomas que não os já previamente incorporados ao sistema.

A utilização do *Festival* continua sendo um ponto positivo, mas acho que se faz necessário o desenvolvimento de outros idiomas, notadamente o idioma português. Uma outra limitação é ter o *Festival* como servidor, o que torna um ponto de investigação ter ele próprio ou uma outra ferramenta TtS (*Text-to-Speech*) executando no lado do cliente junto com a aplicação *talking head* desenvolvida.

Os estudos levaram a constatar que um dos problemas presentes em todos os artigos e que ainda parece estar em aberto na comunidade de animação facial é o casamento entre as expressões faciais e os visemas. Um ponto interessante seria, a partir de um sinal de fala gerado, conseguir extrair os componentes de fala e de emoção, talvez utilizando análise de componentes independentes como em [Cao2003]. Uma outra técnica bastante interessante na separação, ou identificação, dos componentes emocionais e verbais foi apresentada no artigo [Pyun2003]. Nesse caso não seria interessante apenas identificar a importância de cada vértice e definir se ele é mais emocional ou verbal, mas sim, tendo conhecimento dessa importância, no caso dos vértices da região da boca, saber classificá-los como verbais, mas sem perder a contribuição emocional existente neles, ou seja, não perder a emoção na boca (o que hoje acontece no *"Expressive Talking Heads"* [Lucena2002]).

De posse do sistema *talking head* e de uma interface para entrada de textos com marcações de emoção, um ponto interessante é fazer que essa entrada siga algum padrão de especificação. Seguindo essa linha, uma alternativa é o uso da linguagem SSML (*Speech Synthesis Markup Languages*) [W3C2004a], definida pelo W3C[9], o que exigiria o desenvolvimento de um *parser* para essa linguagem (talvez também extensões, pois a linguagem não deve prever a especificação de emoções para a fala). Esse trabalho pode ser desenvolvido em paralelo ao estudo da face, utilizando como fonte para experimentos o sistema "*Expressive Talking Heads"*.

Uma vez investigada a entrada textual, uma outra abordagem seria ter uma interface baseada em fala, como relatou [Chai2003] para seus trabalhos futuros. Para isso poderia

[9] W3C: *World Wide Web Consortium*. URL:*http://www.w3.org/*

se pensar em investigar o padrão VoiceXML [W3C2004b] do W3C para pensarmos como poderíamos, a partir de um sinal de fala, mapear esse sinal numa linguagem padrão para ser sintetizada e passada para o personagem virtual. Essa idéia também levaria à investigação de sistemas de reconhecimento de voz.

Concluindo, existe um interesse particular em investigar o padrão MPEG-4 para animação facial [Pandzic2002] a fim de tentar entender a estrutura que o padrão afirma disponibilizar para o desenvolvimento de animações faciais realistas e, mais precisamente, o desenvolvimento de sistemas *talking heads* em tempo real. Por fim, poderiam ser desenvolvidos experimentos, como os apresentados em [Cunningham2003], a fim de validar quão convincentes e realistas seria o modelo facial desenvolvido com suas expressões faciais foto-realistas.

10. Referências Bibliográficas

[Baker2003] Baker, S.; Matthews, I.; **"Lucas-Kanade 20 years on: A unifying framework part 1: The quantity approximated, the warp update rule, and the gradient descent approximation"**. Publicado no *International Journal of Computer Vision*, 2003.

[Black2004] Black, A.; Clark, R.; **"Festival Speech Synthesis",** Desenvolvido por CSTR – *The Centre for Speech Technology Research, University of Edinburgh*, Versão 2.0, Outubro, 2004. Disponível em http://www.cstr.ed.ac.uk/projects/festival/ (Acesso em 18 Jan. 2005).

[Black1995] Black, M.; Yacoob, Y.; **"Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image otions."**, Publicado nos *Proceedings of International Conference on Computer Vision*, pg. 374-381, 1995.

[Brand1999] Brand, M.; **"Voice Pupperty"**, Publicado nos *Proceedings of ACM SIGGRAPH 1999*, pg. 21-28. ACM Press/Addison-Wesley Publishing Co., 1999.

[Bregler2002] Bregler, C.; Loeb, L.; Chuang, E.; Deshpande, H.; **"Turning to the masters: Motion capturing cartoons"**, Publicado nos *Proceddings of SIGGRAPH 02*, pp. 399-407, 2002.

[Bregler1997] Bregler, C.; Covell, M.; Slaney, M.; **"Video rewrite: driving visual speech with audio."**, Publicado nos *Proceedings of the SIGGRAPH 97 Conference*, pg. 353-360. ACM SIGGRAPH, 1997.

[Buck2000] Buck, I.; Finkelstein, A.; Jacobs, C.; Klein, A.; Salesin, D.; Seims, J.; Szeliski, R.; Toyama, K.; **"Performance-Driven Hand-Drawn Animation"**, Publicado nos *Proccedings of Symposium on Non-Photorealistic Animation and Rendering*, 2000.

[Bui2003] Bui, T.D; Heylen, D.; Nijholt, A. "**Improvements on a simple muscle-based 3D face for realistic facial expressions"**, Publicado nos *Proceedings 16th International Conference on Computer Animation and Social Agents (CASA'2003)*, Rutgers University, New Brunswick, IEEE Computer Society, Los Alamos, CA, ISBN 0-7695-1934-2, 33-40. Disponível em http://wwwhome.cs.utwente.nl/~anijholt/artikelen/casa2003.pdf (Acesso em 13 Dez. 2004).

[Cao2003] Cao, Y.; Faloutsos, P.; Pighin, F.; **"Unsupervised Learning for Speech Motion Editing"**, Publicado nos

Proceedings do 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, San Diego, California, pg. 225 – 231, ISBN ~ ISSN:1727-5288 , 1-58113-659-5, 2003.
Disponível em *http://www.ict.usc.edu/publications/Cao_SCA03.pdf* (Acesso em 05 Jan. 2005).

[Cassell2001] Cassell, J.; Bickmore, T.; Cambell, L.; Vilhjalmsson; Yan, H.; **"More than just a pretty face: conversational protocols and the affordances of embodiment"**, Publicado no *Knowledge-Based Systems*, vol. 14, pp. 22-64, 2001.

[Chai2003] Chai, J.; Xiao, J.; Hodgins, J.; "**Vision-based Control of 3D Facial Animation"**, Publicado nos *Proceedings do 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, San Diego, California, ISBN ~ ISSN:1727-5288 , 1-58113-659-5, 2003.
Disponível em *http://graphics.cs.cmu.edu/projects/face-animation/face-low.pdf* (Acesso em 17 Jan. 05).

[Choe2001] Choe, B.; Lee, H.; Ko, H.; **"Performance-driven muscle-based facial animation"**. Publicado no *Journal of Visualization and Computer Animation*, 12(2), pg. 67-79. Maio, 2001. Disponível em *http://graphics.snu.ac.kr/publications/journals/Choe_2001_JVCA.pdf* (Acesso em 10 Jan. 2005).

[Chuang2002a] Chuang, E.; Deshpande, H.; Bregler, C; **"Facial expression space learning"**, Publicado nos *Proceedings of Pacific Graphics*, 2002.

[Chuang2002b] Chuang, E.; Bregler, C.; **"Performance Driven Facial Animation using Blendshape Interpolation"**, Publicado como *Standford University Computer Science Technical Report*, CS-TR-2002-02, Abril 2002.

[Cunningham2003] Cunningham, D.W.; Breidt, M.; Kleiner, M.; Wallraven, C.; Bülthoff, H.H.; **"How Believable Are Real Faces? Towards a Perceptual Basis for Conversional Animation"**, Publicado nos *Proceedings 16th International Conference on Computer Animation and Social Agents (CASA'2003)*, Rutgers University, New Brunswick, IEEE Computer Society, Los Alamos, CA, ISBN 0-7695-1934-2, 33-40. Disponível em *http://www.hcrc.ed.ac.uk/comic/documents/publications/cunninghamd_believable.pdf* (Acesso em 04 Jan 2005).

[Dutoit1998] Dutoit, T. et al.; **"The MBROLA Project"**, TCTS Lab – *Théorie des Circuits et Traitement du Signal*, *Faculté Polytechinique de Mons*, Bélgica, 1998. Disponível em *http://tcts.fpms.ac.be/synthesis/mbrola.html* (Acesso em 18

Jan. 2005).

[Ekman1971] Ekman, P; **"Universal and cultural differences in facial expressions of emotion"**, Publicado em *Nebraska Symposium on Motoivation 1971*, J. R. Cole, Ed. Lincoln, NE: University of Nebraska Press, 1972, pp. 207-283.

[Essa1997] Essa, I.; Pentland, A.; **"Coding, analysis, interpretation, and recognition of facial expressions"**, Publicado no *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), pg. 757-763, 1997.

[Ezzat2002] Ezzat, T.; Geiger, G.; Poggio, T.; **"Trainable videorealistic speech animation"**; Publicado nos *Proceeding of ACM SIGGRAPH 2002*, pg. 388-398. ACM Press, 2002.

[Gleicher1998] Gleicher, M.; **"Retargetting motion to new Characters"**, Publicado nos *ACM SIGGRAPH 98 Conference Proceedings*, pp. 33-48, 1998.

[Gokturk2001] Gokturk, S. B.; Bouguet, J. Y.; Grzeszczuk. **"A data driven model for monocular face tracking"**, Publicado no *IEEE International Conference on Computer Vision*, pp. 701-708, 2001.

[Guenter1998] Guenter, B.; Grimm, C.; Wood, D.; Malvar, H.; Pighin, F.; **"Making Faces"**. Publicado nos *Proceedings of SIGGRAPH 98. Computer Graphics Proceedings, Annual Conference Series*, pp. 55-66, 1998.

[Huber2001] Huber, D.F.; Hebert, M.; **"Fully automatic registration of multiple 3d data sets"**. Publicado no *IEEE Computer Society Workshop on Computer Vision Beyond the Visible Spectrum (CVBVS 2001)*, 19, pg. 989-1003, Dezembro, 2001.

[Hyvarinen2001] Hyvarinen,, A.; Karhunen, J.; Oja, E. **"Independent Component Analysis"**. Publicado por *John Wiley & Sons*, 2001.

[Jollife1986] Jollife, I. T.; **"Principal Components Analysis"**, Publicado por *New York: Springer*, 1986.

[Joshi2003] Joshi, P.; Tien, W. C.; Desbrun, M.; Pighin, F.; **"Learning Controls for Blend Shape Based Realistic Facial Animation"**, Publicado nos *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation***,** *San Diego, California*, Pg. 187-192, ISBN ~ ISSN:1727-5288 , 1-58113-659-5**,** 2003. Disponível em *http://www.ict.usc.edu/publications/Joshi_SCA03.pdf* (Acesso em 10 Jan. 2005).

[Kalra1992] Kalra, P.; Mangili, A.; Magnenat-Thalmann, N.; Thalmann, D.; **"Simulation of facial muscle actions based on rational free form deformations."** Publicado nos *Proceedings of Eurographics 92*, pg. 59-69, 1992.

[King2000] King, S. A; Parent, R.E.; Olsafsky, B. **"An Anatomically-Based 3D Parametric Lip Model to Support Facial Animation and Synchronized Speech"**, Publicado nos *Proceedings of Deform 2000*, 29-30 Novembro, Genova, pp. 7-19, 2000. Disponível em http://www.cse.ohio-state.edu/research/graphics/research/FacialAnimation/Papers/Deform2000.pdf (Acesso em 03 Jan. 2005).

[Lewis2000] Lewis, J.P.; Cordner, M.; Fong, N.; **"Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation"**. Publicado nos *Proccedings of ACM SIGGRAPH 2000. Computer Graphics Proccedings, Annual Conference Series*, New Orleans, LO, pp. 165-172, 2000.

[Li2001] Li, S.Z.; Gu, L.; **"Real-time multiview face detection, tracking, pose estimation, alignment, and recognition"**. Publicado no *IEEE Conference on Computer Vision and Pattern Recognition Demo Summary*, 2001.

[Lien1998] Lien, J.; Cohn, J.; Kanade, T.; Li, C.C.; **"Automatic facial expression recognition based on FACS action units."** Publicado nos *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pg. 390-395, 1998.

[Liu2001] Liu, Z.; Shan, Y.; Zhang, Z.; **"Expressive expression mapping with ratio images"**. Publicado no *Computer Graphics Annual Conferences Series, SIGGRAPH*, Agosto de 2001, Pg. 271-276. Slides da apresentação disponíveis em: http://online.cs.nps.navy.mil/DistanceEducation/online.siggraph.org/2001/Papers/07_AnimationAndExpression/liu.pdf (Acesso em 22 Dez 2004).

[Lucena2002] Lucena, P.S.; **"Expressive Talking Heads: Um Estudo de Fala e Expressão Facial em Personagens Virtuais"**, Dissertação de Mestrado, Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Junho, 2002. Disponível em http://www.telemidia.puc-rio.br/~pslr/publicacoes/tese.pdf (Acesso em 15 Fev 2005).

[Noh2001] Noh, J.Y.; Neumann, U. **"Expression Cloning"**. Publicado no *Computer Graphics, Annual Conference Series, SIGGRAPH*, Agosto de 2001, Pg. 277-288.

[Pandzic2002] Pandzic, I.S.; Forchheimer, R.; **"MPEG-4 Facial Animation: The Standard, Implementation and Applications"**, Publicado por *John Wiley & Sons Ltd.*, ISBN 0-470-84465-5, 2002.

[Parke1996] Parke, F.I.; Waters, K. **"Computer Facial Animation"**, Publicado por *AK Peters*, ISBN 1-56881-014-8, 1996.

[Parke1982] Parke, F. I.; **"Parameterized models for facial animation"**, Publicado no *IEEE Computer Graphics and Applications*, Vol. 2 No. 9, pp. 61-68, 1982.

[Parke1974] Parke, F.I.; **"A parametric model for human faces"**; *PhD thesis*, Universidade de Utah, Salt Lake City, Utah, UTEC-CSc-75-047, Dezembro de 1974.

[Parke1972] Parke, F.I.; **"Computer generated animation of faces"**, *Master's thesis*, Universidade de Utah, 1972.

[Perlin2003] Perlin, K.; **"Kinetic Sculpture (or The Clay Become Flesh)".** Disponível em http://mrl.nyu.edu/~perlin/experiments/head/ (Acesso em 17 Jan. 2005)

[Pighin1998] Pighin, F.; Hecker, J.; Lischinski, D.; Szeliski, R.; Salesin, D.H.; **"Synthesizing realistic facial expressions from photographs"**. Publicado no *Computer Graphics, Annual Conference Series, SIGGRAPH*, Julho de 1998, Págs. 75-84. Disponível em: http://grail.cs.washington.edu/pub/papers/Pighin98.pdf (Acesso em 22 Dez 2004).

[Platt1981] Platt, S. M.; Badler, N. I.; **"Animating facial expressions."**; Publicado no *Computer Graphics*, 15(3), pg. 245-252, 1981.

[Pyun2003] Pyun, H.; Kim, Y.; Chae, W.; Kang, H. W.; Shin, S. Y.; **"An Example-Based Approach for Facial Expression Cloning"**, Publicado no *Symposium on Computer Animation, Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer* Animation, San Diego, California, pg. 167-176, ISBN ~ ISSN:1727-5288 , 1-58113-659-5, 2003. Disponível em http://cg.kaist.ac.kr/Papers/Pyun_136.pdf (Acesso em 12 Jan. 005).

[Shreiner2003] Shreiner, D.; Woo, M.; Needer, V.; Davis, T.; **"OpenGL Programming Guide: The Oficial Guide to Learning OpenGL"**, Publicado por *Addison-Wesley Pub Co, Version 1.4, fourth Edition*, ISBN: 0321173481, Novembro, 2003.

[Terzopoulos1993] Terzopoulos, D.; Waters, K.; **"Analysis and synthesis of facial image sequences using physical and anatomical models"**, Publicado no *IEEE Trans. On Pattern Analysis and Machine Interlligence*, 15, pp. 569-579, 1993.

[Thalmann1988] Magnenat-Thalmann, N.; Primeau, N. E.; Thalmann, D.; **"Abstract muscle actions procedures for human faces."** Publicado no *Visual Computer* 3(5), pg. 290-297, 1988.

[W3C2004a] W3C Consortium; **"SSML – Speech Synthesis Markup Language Especification – Version 1.0"**, *W3C Recommendation,* Setembro, 2004. Disponível em

http://www.w3.org/TR/speech-synthesis/ (Acesso em 18 Jan. 2005).

[W3C2004b] W3C Consortium; **"VoiceXML – Voice Extensible Markup Language Especification – Version 2.0"**, *W3C Recommendation,* Março, 2004. Disponível em http://www.w3.org/TR/voicexml20/ (Acesso em 18 Jan. 2005).

[Waters1987] Waters, K. **"A muscle model for animating three-dimensional facial expressions"**, Publicado no *Computer Graphics (SIGGRAPH'87)*, 21(4), Julho, 17-24. Informação em http://www.nbb.cornell.edu/neurobio/land/OldStudentProjects/cs490-95to96/hjkim/waters87.html (Acesso em 16 Dez. 2004).

[Williams1990] Williams, L.; **"Performance driven facial animation"**, Publicado no *Proceddings of SIGGRAPH 90*, pp. 235-242, 1990.

[Ye1997] Ye, Y. **"Interior Point Algorithms: Theory and Analysis"**. Publicado por *John Wiley*, 1997.

[Zhang2003] Zhang, Q; Liu, Z; Guo, B; Shum, H. **"Geometry-Driven Photorealistc Facial Expression Synthesis"**, Publicado nos *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, San Diego, California, ISBN/ISSN: 1727-5288, 1-58113-659-5, Pg. 177-189. Disponível em http://www.research.microsoft.com/asia/dload_files/group/ig/2003/sca03.pdf (Acesso em 17 Fev. 2005).