PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Aprendizado e Utilização do Estilo de Movimento Facial na Animação de Avatares

Adriana Braun

Tese apresentada como requisito à obtenção do grau de Doutor em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Profª Drª Soraia Raupp Musse

Porto Alegre
2014

Dados Internacionais de Catalogação na Publicação (CIP)

B825a Braun, Adriana
Aprendizado e utilização do estilo de movimento facial na animação de avatares / Adriana Braun. – Porto Alegre, 2014.
124 p.

Tese (Doutorado) – Fac. de Informática, PUCRS.
Orientador: Profª. Drª. Soraia Raupp Musse.

1. Informática. 2. Animação por Computador. 3. Computação Gráfica. 4. Avatares. 5. Redes Neurais (Computação). I. Musse, Soraia Raupp. II. Título.

CDD 006.6

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "Aprendizado e Utilização do Estilo de Movimento Facial na Animação de Avatares", apresentada por Adriana Braun, como parte dos requisitos para obtenção do grau de Doutora em Ciência da Computação, aprovada em 08/08/2014 pela Comissão Examinadora:

Profa. Dra. Soraia Raupp Musse — PPGCC/PUCRS
Orientadora

Profa. Dra. Isabel Harb Manssour — PPGCC/PUCRS

Prof. Dr. Marcelo Walter — UFRGS

Prof. Dr. Bruno Feijó — PUCRIO

Homologada em...../...../....., conforme Ata No. ..... pela Comissão Coordenadora.

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 - P. 32 - sala 507 - CEP: 90619-900
Fone: (51) 3320-3611 - Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

AGRADECIMENTOS

Esse trabalho não foi construído por um indivíduo. Ao longo do processo, muitas pessoas contribuíram para o seu desenvolvimento. Quero aqui expressar meus profundos agradecimentos a essas pessoas.

Primeiramente, agradeço à minha orientadora, professora Dra. Soraia Raupp Musse pela inspiração, auxílio e valiosos conselhos. Sua energia, determinação e criatividade são ímpares e contagiam seus alunos. Muito obrigada pelo apoio incondicional ao longo desses anos.

Agradeço ao meu marido, Eduardo, pelo incentivo e suporte. Seu carinho, apoio e paciência me dão segurança e motivação para enfrentar qualquer desafio. Obrigada por estar a meu lado e compartilhar sua vida comigo.

Aos meus pais e irmãos por fazerem parte de minha vida e por serem tão especiais.

Aos colegas do VHLab pelo companheirismo e pela ótima convivência e espírito de equipe. Em especial, gostaria de agradecer ao Leandro Dihl, colega, amigo e colaborador desde o início do período de doutorado. À Rossana Baptista Queiroz, por compartilhar comigo os desafios da área de animação facial e por adaptar seu trabalho e colocá-lo à disposição para visualização dos resultados desta tese. Agradeço também ao Vinicíus Cassol pela alegria e companheirismo.

À professora Dra. WonSook Lee pela acolhida na Universidade de Ottawa e pela atenção e ajuda para o desenvolvimento desse trabalho. Agradeço também aos colegas de laboratório, que foram grandes amigos e colaboradores durante minha estada no Canadá. Em especial agradeço à Niloofar Aghayan e Alberto Chaves pela colaboração no trabalho. Também não poderia deixar de citar Ava Ahadipour, Sahar Aghayan e Iman Eshragui pela valiosa amizade.

Agradecimentos também à Pró-reitoria de Pesquisa da PUCRS pelo apoio financeiro durante o último semestre do curso e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de doutorado sanduíche PDSE, realizado na Universidade de Ottawa, Canadá.

Esse trabalho foi desenvolvido com apoio financeiro da Hewlett-Packard Brasil Ltda, usando incentivos da Legislação Brasileira(Lei nº 8.248 de 1991).

Aprendizado e Utilização do Estilo de Movimento Facial na Animação de Avatares

RESUMO

Esse trabalho apresenta uma metodologia, denominada Persona, para aprendizado e transferência do estilo de movimento facial de atores para a animação de avatares. Por meio dessa metodologia, pode-se guiar a animação das faces de avatares com o estilo de movimento de um ator específico, através da atuação de usuários quaisquer. Dessa forma, o avatar poderá expressar os movimentos faciais que o usuário executa, porém replicando as particularidades dos movimentos faciais do ator, por meio da utilização da Persona. Para construção do estilo de movimento facial dos atores, utilizou-se como dados de entrada pontos da face obtidos por rastreamento em sequências de imagens e informações presentes em bancos de dados de expressões faciais tridimensionais, anotadas de acordo com o Sistema de Codificação Ações Facial (FACS). Esses dados foram submetidos à análise de componentes principais e, então, utilizados para treinamento de redes neurais artificiais. Com esses classificadores podem-se reconhecer automaticamente as unidades de ação na expressão do usuário e encontrar os parâmetros equivalentes no estilo de movimento do ator. O resultado do processo é o fornecimento desses parâmetros para sistemas de animação. O protótipo desenvolvido como prova de conceito foi utilizado em casos de estudo, cujos resultados são apresentados. Indicações de trabalhos futuros também serão discutidas.

Palavras-chave: Animação facial, Estilo de Movimento, Redes Neurais Artificiais, Reconhecimento de expressões faciais.

Learning and Using Facial Motion Style for Avatar Animation

ABSTRACT

This work presents a methodology, named Persona, for learning and transfer of facial motion style of an actor in order to provide parameters for avatar facial animation. Through this methodology, we can drive the facial animation of avatars using the motion style of a particular actor, through the performance of any user. Thus, the avatar can express the facial movements that the user is performing, but replicating the particularities of the actor's facial movements, based on his or her Persona. In order to build the facial motion style model of an actor, we used points tracked on image sequences of the actor performance as input data. We also used a database of three-dimensional facial expressions, annotated according to the Facial Action Coding System (FACS). Principal components analysis was performed using these data. Afterwards, artificial neural networks were trained to recognize action units both in the actor and user's performances. Given these classifiers, we can automatically recognize action units in the user's expression and find the equivalent parameters in the actor's motion style model. The result of the process is the provision of such parameters to facial animation systems. The prototype developed as proof of concept has been used in case studies, whose results are discussed. Future work are also addressed.

Keywords: Facial animation, Motion Style, Artificial Neural Networks, Facial Expression Recognition.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

PDA	*Performance Driven Animation*
SVM	*Support Vector Machines*-Máquinas de Suporte Vetorial
AU	Action Units
FACS	Facial Action Coding System
ASM	Active Shape Models
AAM	Active Apearance Models
3DMM	3D Morphable Models
LKT	Lucas–Kanade–Tomasi
PCA	*Principal Component Analysis*
PC	*Principal Component*
RNA	Rede Neural Artificial
MLP	*Multilayer Perceptron*
FP	*Feature Points*
RGBD	*Red, Green, Blue and Depth*
DHM	Distância de Hausdorff Modificada

SUMÁRIO

1. INTRODUÇÃO

"O rosto é o retrato da alma". Essa citação do filósofo romano Marcus Tullius Cicero (106A.C a 46A.C.), numa livre interpretação, apresenta o rosto e as expressões faciais como a manifestação do estado interior de uma pessoa. Através da face e de seus movimentos, os seres humanos expressam voluntária ou involuntariamente sentimentos, traços culturais, intenções e até mesmo manifestações fisiológicas como alívio ou dor. É através da face que mostramos o que se passa em nossa mente, o que a configura como uma interface de comunicação com outros seres. A mesma frase dita com o acompanhamento de um sorriso ou de um piscar de olhos pode mudar completamente o sentido do que está sendo dito.

Segundo Mehrabian [46], apenas 7% da informação sobre os sentimentos e atitudes de quem fala são transmitidos através das palavras proferidas. Quanto ao restante, 38% da informação provêm de sinais paralinguísticos (forma ou tom com que as palavras são ditas) e 55% são transmitidos pela postura e ações, especialmente pelas expressões faciais. Apesar deste modelo ser contestado por alguns psicólogos que sustentam que essas proporções não são válidas em todas as circunstâncias, não há contestação sobre a importância dos sinais não verbais na comunicação humana, mesmo que as porcentagens oscilem em torno desses valores.

Dado esse quadro, é natural que o interesse sobre o estudo do mapeamento de expressões faciais para modelos computacionais de avatares humanos seja cada vez maior. Nos últimos anos, a academia e as indústrias cinematográfica e de jogos têm conseguido impressionantes avanços nesse campo de pesquisa, tornando os avatares e agentes comunicativos mais semelhantes aos humanos, no que diz respeito à expressão facial. Tais avanços tornaram possível que gestos e expressões faciais de um ator possam ser mapeados para um modelo tridimensional, de tal forma que os movimentos do ator sejam convertidos na animação do personagem virtual com extrema precisão. Hoje, esse processo é feito usando marcadores no rosto da pessoa (maquiagem ou pontos reflexivos) que são seguidos por algoritmos de visão computacional e convertidos em arquivos de animação. O desempenho do ator, nesses casos, é capturado por equipamentos especiais, que fornecem múltiplas imagens do ator em alta resolução. Esse processo é denominado "Captura de Movimento" (*Motion Capture* [56]). Por exemplo, A Figura 1.1 mostra a preparação da atriz Zoë Saldana para interpretar a personagem "Neytiri" em uma cena do filme "Avatar"[1]. À direita, pode-se ver o seu desempenho mapeado para o modelo computacional da personagem.

Para esse tipo de aplicação, a reflexão da expressão do ator no avatar não precisa necessariamente ocorrer em tempo real, ou seja, os algoritmos de análise do movimento podem ser lentos e os dados obtidos podem ser pós-processados por artistas para garantir uma animação precisa e convincente. Em outras palavras, aplicações com esse nível de qualidade requerem um tempo considerável para processamento com muito investimento em pós-processamento, o que inviabiliza aplicações dessas

[1] Avatar: Fox Movies, 2009. Dirigido por James Cameron.

técnicas, quando se necessita que a animação do avatar seja feita em tempo real [31]. Além disso, o custo desses sistemas é elevado, de forma que, em geral, não são utilizadas em aplicações de uso doméstico.

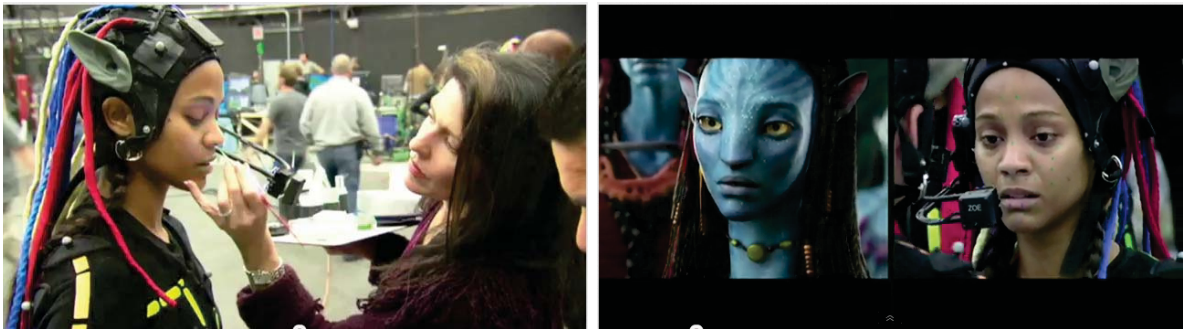


Figura 1.1: Imagens do filme Avatar. À direita, a atriz Zoë Saldana está sendo preparada com marcadores em seu rosto para ter seu desempenho capturado. À esquerda, o desempenho da atriz é mapeado para o personagem 3D.

Por outro lado, em aplicações de uso mais popular, é desejável que as informações de entrada sejam provenientes de dispositivos de aquisição de baixo custo, sem que haja necessidade de utilização de marcadores por parte do usuário [83]. Nessas aplicações, as técnicas computacionais envolvidas se enquadram na área denominada de Animação Dirigida por Performance (*Performance Driven Animation - PDA*), que será melhor descrita na Seção 2.1. Em geral, nesses casos, a qualidade da animação obtida não é tão boa e não são notadas nuances ou detalhes que identifiquem as particularidades do ator cujas expressões estão sendo rastreadas e mapeadas. O estudo do mapeamento desses detalhes específicos é ainda um campo de pesquisa em aberto, no qual esse trabalho pretende contribuir. Especificamente, esse trabalho pretende apresentar uma técnica para aprendizado e mapeamento dessas particularidades.

Uma das motivações do presente trabalho é o fato de que as individualidades que as pessoas demonstram por meio de suas expressões faciais são fundamentais na definição de sua identidade ou na identidade de personagens que elas interpretam. Bons atores são capazes de representar diferentes personagens adotando posturas e expressões típicas de cada um. Jim Carey, Jack Nicholson e Al Pacino, por exemplo, mudam fisicamente na caracterização de cada um deles e a alteração dos padrões de expressão facial fazem parte importante desse repertório de mudanças. A Figura 1.2 mostra imagens do ator Jim Carey, que é conhecido por suas expressões faciais típicas e exageradas, encenando três diferentes personagens com padrões de expressão facial muito marcantes e distintos. Na imagem da esquerda, ele aparece interpretando "O Máscara"[2]. No centro, ele aparece com expressões bem mais comedidas no filme "Brilho Eterno de uma Mente sem Lembranças"[3]. Em outro exemplo, à direita, ele aparece como o ingênuo Truman em "O Show de Truman"[4]. Nesses três casos, o ator construiu seus personagens de forma que eles se expressem de maneira distinta. Cada

[2] `http://www.imdb.com/title/tt0110475/?ref\_=fn\_al\_tt\_4`, acessado em 10/07/2014
[3] `http://www.imdb.com/title/tt0338013/?ref\_=fn\_al\_tt\_1`, acessado em 10/07/2014
[4] `http://www.imdb.com/title/tt0120382/?ref\_=fn\_al\_tt\_1`, acessado em 10/07/2014

um deles se constitui numa "persona" própria com expressões, trejeitos e posturas diferenciadas. Por exemplo, não se espera que o ator se utilize do mesmo repertório de expressões ao dizer a mesma frase interpretando o "Máscara" e interpretando "Truman". Segundo Piangiani [58] a palavra italiana "persona" deriva do latim *per sonare* ou em português, "soar através" ou "ressoar". Originalmente, era o nome dado a um tipo de máscara que atores utilizavam no teatro antigo. Essa máscara fazia com que a voz do ator ressoasse e fosse amplificada. Contribuía ainda para dar a aparência que o papel exigia. A partir da Roma antiga, o significado mudou para indicar o personagem em uma interpretação teatral.

Dada a importância da expressão facial, é de interesse que um sistema computacional seja capaz de aprender as particularidades de cada conjunto de expressões que, quando mapeadas para um avatar, permitam melhor representar cada um dos personagens. Assim, em um momento futuro, o avatar poderá executar as mesmas expressões características, sem a necessidade de nova captura ou análise do movimento do ator. Poder-se-ia, por exemplo, pensar em um jogo cujo personagem seja o "Máscara" e cuja animação seja dirigida pela performance de um jogador. Se o estilo de movimento do "Máscara" for convenientemente aprendido, o avatar do Máscara poderá ser guiado pelas expressões do jogador, mas preservar o jeito de se expressar que o ator Jim Carey executou quando interpretou o personagem.

Figura 1.2: Imagens do ator Jim Carey, encenando três personagens com estilos de movimento facial distintos.

Da mesma forma que atores podem mudar a forma de se expressar, cada pessoa adota uma postura diferente em diferentes circunstâncias. Não nos portamos da mesma maneira em um jantar informal, dando uma palestra para pessoas ilustres ou numa entrevista de emprego. Pode-se dizer que, para cada uma dessas situações, adotamos uma "persona" diferente e o conjunto dessas diferentes "personas" podem definir os vários papéis que podemos exercer socialmente. As expressões faciais que uma pessoa adota em cada uma dessas situações são parte importante desse conjunto de atitudes.

No escopo desse trabalho, denominou-se Persona a uma estrutura de dados que sintetiza o conjunto de expressões faciais que um indivíduo apresenta como parte do seu estilo de movimento numa ou em mais circunstâncias. Assim sendo, a Persona do ator Jack Nicholson será diferente ao interpretar o portador de transtorno obsessivo compulsivo Melvin Udall no filme Melhor é Impos-

sível[5] do que ao interpretar o Coringa no filme Batman[6] que difere ainda do Jack Torrance em O Iluminado[7]. Seu jeito de sorrir, o movimento das sobrancelhas, mesmo a manifestação do estado de humor interno são diferentes para cada personagem. Essas peculiaridades são relevantes e devem ser preservadas quando se deseja mapear tais expressões para um avatar.

Nos últimos anos, a área de pesquisa de animação computadorizada tem voltado sua atenção para o aprendizado e a reprodução das particularidades do movimento de atores nos modelos geométricos computacionais [6, 20, 47, 50, 78, 82]. Tem se tornado frequente na literatura a menção ao aprendizado automático do *Estilo de Movimento* de pessoas cujo desempenho tenha sido capturado e digitalizado. Segundo Torresani e colegas [78], a análise do movimento humano pode ser pensada como a interação de dois fatores, tradicionalmente denominados conteúdo e estilo. O conteúdo geralmente refere-se à natureza da ação no movimento (por exemplo, caminhar, saltar, sentar), enquanto o estilo denota a forma particular que a ação é realizada. Conforme poderá ser verificado na Seção 2.3, porém, a maior parte dos estudos realizados refere-se ao aprendizado e utilização do estilo de movimento de ações como caminhada, corrida e execução de atividades esportivas. Poucos trabalhos têm sido dedicados ao aprendizado do estilo de movimento da face para aplicações em tempo real.

No contexto do presente trabalho, a natureza da ação será determinada pela presença de unidades de ação para o movimento aparente da face, de acordo com o Sistema de Coodificação de Ações Faciais (FACS) proposto por Ekman e colaboradores em [18] e apresentadas na Seção 3.1. O estilo será a forma particular com que essas unidades de ação são realizadas pelo ator.

O problema de pesquisa desse trabalho é a investigação sobre a possibilidade de aprendizado do estilo de movimento da face de atores para utilização em avatares. Além disso, propõe-se uma técnica que visa utilizar esse estilo de movimento facial - ou Persona - em avatares guiados pela performance de usuários que podem ou não corresponder ao ator que originou a Persona. Deve-se salientar que, no escopo do presente trabalho, o termo ator não necessariamente se refere a profissionais da área. Qualquer pessoa cujo estilo de movimento facial se queira utilizar posteriormente para animação de um avatar se enquadra nessa definição. A próxima seção detalha os objetivos desse trabalho.

1.1 Objetivos

1.1.1 Objetivo geral

Dada a atual tecnologia, esse trabalho se propõe a investigar se, com a informação disponível por algoritmos de rastreamento de componentes faciais do estado da arte, é possível aprender e codificar a Persona de um determinado ator A. Além disso, pretende-se transferir esse estilo de movimento para a face de um avatar, que será animado de forma a expressar a Persona de A. Finalmente, deseja-se que a animação do avatar de A seja dirigida pela performance de outro indivíduo usuário

[5] `http://www.imdb.com/title/tt0119822/`
[6] http://www.imdb.com/title/tt0096895/?ref_=nm_flmg_act_18
[7] `http://www.imdb.com/title/tt0081505/?ref\_=nm\_knf\_t1`

U do sistema, porém, com o estilo de movimento de A. Assim sendo, quando por exemplo o usuário U sorrir, o avatar deverá sorrir também. Entretanto, esse sorriso deverá ser semelhante ao sorriso de A e não de U.

1.1.2 Objetivos específicos

- Propor uma estratégia para aprendizado do estilo de movimento facial (Persona) de um indivíduo A, cujas expressões faciais foram rastreadas automaticamente em imagens adquiridas por uma câmera de vídeo de uso doméstico;

- Prover parâmetros para um sistema de animação de um avatar de A com o estilo de movimento de A guiado pela performance de um usuário U, cujo movimento facial é capturado por câmera monocular;

- Desenvolver um protótipo;

- Avaliar os resultados obtidos.

1.2 Visão Geral do Modelo

A Figura 1.3 mostra uma visão geral da estratégia proposta para que sejam atingidos os objetivos listados na seção anterior. Essa estratégia tem como base a construção e utilização da estrutura de dados Persona, respectivamente à esquerda e à direita da Figura 1.3. De forma geral, essa estrutura de dados corresponde a um conjunto de parâmetros de animação obtidos a partir de pontos da face do ator rastreados em sequências de vídeo e de informações providas por um banco de dados de faces humanas. Esses parâmetros devem ser classificados de acordo com as ações presentes na atuação do ator, de forma que seja associada a eles uma interpretação em alto nível que torne possível sua identificação posterior. O esquema da esquerda da Figura 1.3 mostra resumidamente esse processo de construção da Persona.

Decidiu-se utilizar apenas informações de vídeo na construção da Persona em virtude da maior facilidade de aquisição de dados e da possibilidade de se trabalhar com informações de atores que não estarão presentes para a captura de seu movimento por equipamentos mais sofisticados. Assim, pode-se, inclusive, pensar na construção da Persona de atores falecidos para utilização de seu estilo de movimento em situações que não tenham sido gravadas enquanto estavam vivos. Por exemplo, alguém poderia realizar um comercial utilizando um avatar de Marilyn Monroe, cujas expressões faciais seriam semelhantes àquelas que eram características de seus personagens, dizendo um texto que ela nunca tenha proferido e em circunstâncias distintas daquelas em que ela foi filmada.

As informações de entrada do sistema são, portanto, constituídas de pontos rastreados na face do ator, que devem ser obtidos por meio de ferramentas de análise da face que possam prover esses dados em tempo real e com boa robustez. É essencial que se tenha o contorno dos lábios,

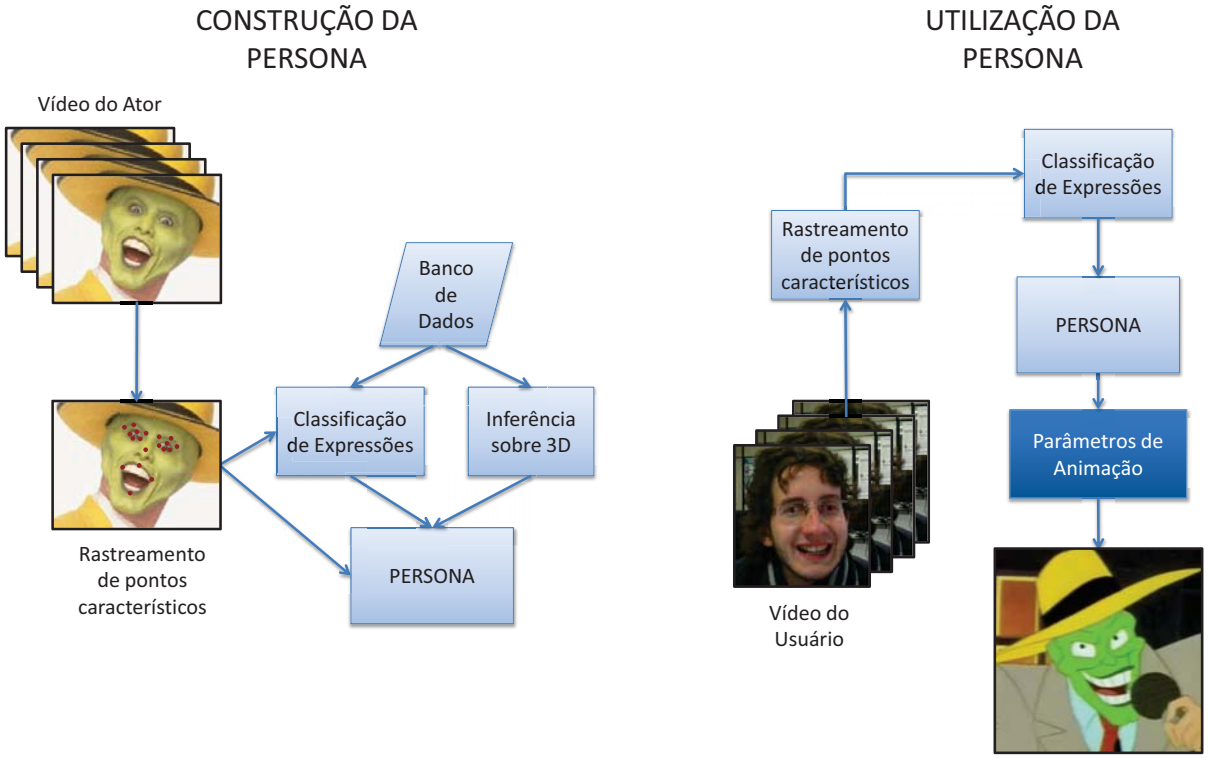


Figura 1.3: Visão geral do modelo proposto.

olhos e sobrancelhas com boa precisão para que sutilezas presentes na atuação do ator possam ser incorporadas à sua Persona.

O banco de dados utilizado para auxiliar na construção da Persona é composto por nuvens de pontos digitalizadas e anotadas de acordo com as ações presentes nas expressões das pessoas. Dadas essas informações, foram treinados classificadores capazes de identificar essas ações em imagens de faces. Além disso, o banco de dados é utilizado para inferência do deslocamento de componentes faciais na direção perpendicular ao plano da face. Dessa forma, movimentos dos lábios "para frente" ou "para trás" do plano da imagem podem ser estimados, mesmo que essas informações não estejam explícitas nos dados rastreados nos vídeos de entrada. As ações do ator em cada quadro do vídeo de entrada são classificadas de acordo com o treinamento e são determinados os parâmetros de animação associados aos pontos rastreados.

Uma vez construída a Persona do ator, ela será utilizada para prover parâmetros de animação para o avatar. Para que esses parâmetros sejam obtidos, os dados provenientes de rastreamento de pontos característicos da face de um usuário (identificado no esquema à direita da Figura 1.3) são analisados, as ações presentes são classificadas e os parâmetros contidos na Persona do ator correspondentes a essas ações são identificadas. Assim, a atuação do usuário irá guiar a animação do avatar, mas com parâmetros de animação provenientes da Persona do ator.

A próxima seção apresenta a organização dessa tese.

1.3 Organização da Tese

O próximo capítulo é destinado a uma breve revisão do estado da arte em três áreas distintas correlatas a esse trabalho: animação facial dirigida por performance, aprendizagem e utilização do estilo de movimento de atores cuja performance tenha sido capturada e técnicas de reconhecimento de expressões faciais.

O Capítulo 3 apresenta conceitos fundamentais e técnicas utilizadas no desenvolvimento do modelo computacional proposto. Entre esses conceitos, estão o sistema de codificação de ações faciais (FACS), métodos de rastreamento de pontos característicos da face em tempo real e métodos de reconhecimento de padrões para identificar os movimentos da face humana. Além disso, serão descritos parâmetros utilizados para guiar a animação facial e o banco de dados de faces digitalizadas utilizado nesse trabalho. São apresentadas também técnicas de alinhamento e comparação entre nuvens de pontos.

O modelo computacional desenvolvido para solução do problema apresentado será descrito em detalhes no Capítulo 4. O Capítulo 5 apresenta resultados obtidos e, finalmente, o Capítulo 6 discute considerações finais e indica trabalhos futuros a serem realizados.

2. ESTADO DA ARTE

Este capítulo apresenta uma breve revisão do estado da arte em três áreas correlatas a esse trabalho: Animação Dirigida por Performance, reconhecimento automático de expressões faciais e aprendizado e utilização do estilo de movimento.

Trabalhos na área de pesquisa de Animação Dirigida por Performance são discutidos pois a animação dos avatares no modelo proposto nesta tese deve seguir as ações faciais de um usuário frente à uma câmera de vídeo. Essas ações devem ser identificadas por técnicas de reconhecimento de padrões e, por isso, torna-se relevante uma revisão bibliográfica deste campo de pesquisa (Seção 2.2). Com base na identificação feita, o sistema deve selecionar uma ação correspondente à ação do usuário no estilo de movimento do ator. Por isso, trabalhos que se propõem ao aprendizado de estilos de movimento serão citados na Seção 2.3.

2.1 Animação Dirigida por Performance

A tarefa de animar a face de um avatar a partir da informação da performance de um ator adquirida por câmeras monoculares tem recebido considerável atenção dos pesquisadores nos últimos anos. Entretanto, esse ainda é um tópico de pesquisa em aberto pois, para animar o modelo 3D da face convincentemente e em tempo real, são requeridos parâmetros de animação estáveis, com um bom nível de precisão e que sejam obtidos rapidamente.

Os algoritmos de rastreamento de pontos chave da face, apesar dos bons resultados mostrados na Seção 3.2, são sujeitos a falhas. Por isso, quando se utilizam essas técnicas para Animação Dirigida por Performance (*Performance Driven Animation*), não se pode simplesmente mapear os dados brutos para o avatar. Muitas vezes são necessárias correções nos dados rastreados de forma a garantir a qualidade dos parâmetros de animação gerados. Para realizar essas correções, estratégias para estabilização e interpretação de resultados obtidos por algoritmos de visão computacional são adicionadas pelos pesquisadores da área como uma camada intermediária entre os dados brutos rastreados e a geração de parâmetros de animação. Essa seção apresenta algumas dessas estratégias.

Chai et al. [87] descreveram uma abordagem para animação de avatares baseada em exemplos. Nessa abordagem, um grupo relativamente pequeno de parâmetros de controle são extraídos de vídeo e combinados com o conhecimento adquirido previamente em arquivos de informações de captura de movimentos. Uma das contribuições do trabalho foi propor um método para associar o movimento dos poucos pontos controle da imagem do vídeo a um movimento de alta qualidade extraído dos arquivos de captura de movimentos pré-processados e armazenados em um banco de dados. Uma desvantagem do modelo proposto é que, como inicialização, é necessário que o usuário selecione na imagem de sua face dezenove pontos de controle que, a partir de então, serão rastreados. Resultados obtidos por essa abordagem são mostrados na Figura 2.1.

Stoiber et al. apresentaram em [71] um sistema para combinar a animação de avatares baseada

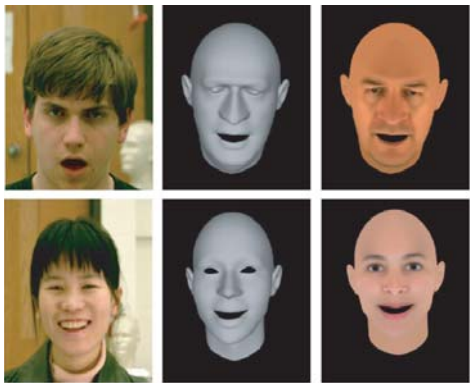


Figura 2.1: Resultados obtidos por Chai et al. [87].

em parâmetros (controlada por um animador) com a animação baseada em performance. O sistema é baseado na construção de espaços de aparência obtidos por modelos de aparência ativos (explicados na Seção 3.2) aplicado em bancos de dados. O sistema foi utilizado para a animação de avatares semelhantes a humanos e para reuso em avatares não semelhantes a humanos, tanto bidimensionais quanto tridimensionais. O sistema é restrito a mapeamento de emoções (não foi desenvolvido para unidades de ação nem visemas). O modelo paramétrico foi mapeado para uma esfera 3D, de forma que o animador pode controlar a animação facial do avatar selecionando um ponto dessa esfera.

Saragih [63] e colaboradores propuseram um sistema para animação facial de fantoches virtuais combinando um modelo genérico de transição de expressões com exemplos individualizados gerados sinteticamente. O modelo genérico utiliza como base o padrão MPEG-4 (descrito na Seção 3.4.1) e é construído a partir de um banco de dados de cerca de 200 imagens anotadas, contendo expressões faciais de diferentes usuários. Nesse treinamento, são aprendidas tanto as expressões alvo como funções de mapeamento da expressão neutra para cada uma delas. Segundo os autores, o uso puro e simples desse modelo pode levar à restrições na articulação do avatar ou deformações se o usuário tiver uma estrutura facial diferente do modelo genérico aprendido. Por isso, as funções de mapeamento aprendidas são utilizadas para construir sinteticamente expressões alvo personalizadas do usuário a partir de uma única imagem de entrada. Além disso, as funções de mapeamento são usadas para gerar automaticamente expressões alvo na face do avatar. Dadas as expressões alvo sintéticas do usuário e as respectivas expressões alvo sintéticas do avatar, um novo mapeamento é aprendido e utilizado em tempo de execução. Durante a execução do sistema, o rastreamento dos pontos chave na sequência de vídeo é feito utilizando um algoritmo de alinhamento facial não rígido que provê tanto a forma quanto a textura da face do usuário que são mapeados para o avatar. Resultados podem ser observados na Figura 2.2. Os resultados apresentados mostram uma alta correlação entre a expressão do usuário e a do avatar.

Rhee et al. [60] propuseram um sistema para animação dirigida por performance a partir de informações capturadas por câmeras monoculares de baixo custo como câmeras *web*. Os autores utilizaram estratégias distintas para a parte superior e inferior da face. Para a parte superior, o rastreamento é baseado em características locais e restringido por propagação de crença (*Belief Propagation*). Um detector de estados para os olhos foi desenvolvido para reconhecimento do rápido

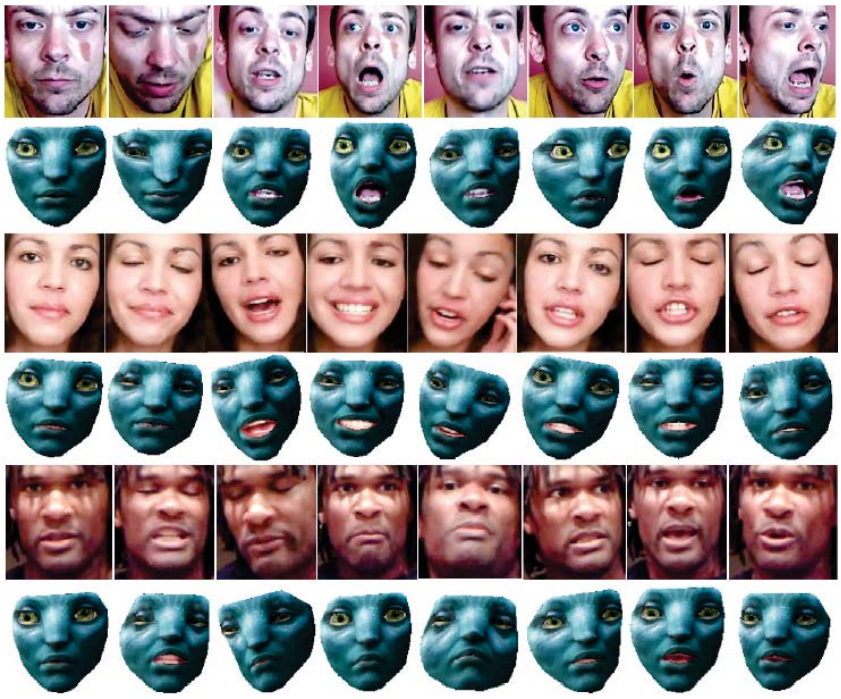


Figura 2.2: Resultados obtidos por Saragih e colaboradores em [63].

movimento de piscar de olhos. Para a parte inferior, foi desenvolvido um algoritmo para casamento da aparência global com exemplos, de forma que os dados extraídos da imagem 2D são comparados com um espaço de onze expressões básicas. Uma técnica de deformação baseada em exemplos lida com os detalhes dinâmicos locais no avatar que não são capturados diretamente do vídeo. Dessa forma, os parâmetros de alta dimensionalidade requeridos para uma boa animação facial são obtidos a partir dos dados esparsos provenientes da análise da imagem 2D. A Figura 2.3 mostra resultados obtidos pelos autores. Nota-se uma boa qualidade de animações, mas as expressões mostradas não representam casos desafiadores, pois tratam-se de expressões básicas sem movimentos extremos.

Figura 2.3: Resultados obtidos por Rhee e colaboradores em [60]. Nas imagens superiores, são mostrados os pontos rastreados nas imagens dos usuários; nas imagens inferiores, os pontos representam os pontos de controle para a animação.

Alguns autores têm se concentrado na captura do movimento dos lábios durante a fala. Pei e Zha [57], por exemplo, desenvolveram um sistema no qual há transferência do movimento dos lábios do sujeito filmado para um modelo 3D. Esse sistema depende do aprendizado de configurações faciais

chave selecionadas para atender cada vogal e consoante do idioma mandarim. Esse aprendizado ocorre por treinamento por meio de sequências de vídeo, utilizando classificação K-means. Nesse trabalho, o modelo 3D é adquirido através da utilização de scanners 3D.

Zhang et al. [90] propuseram uma metodologia para PDA baseando-se no padrão MPEG-4 (descrito na Seção 3.4.1). Os autores propuseram uma abordagem probabilística para integração das unidades de ação do Sistema de Codificação de Ações Faciais (descrito na Seção 3.1) com os parâmetros de animação requeridos pelo padrão MPEG-4. A detecção de pontos chave nas imagens das faces é feita por meio da comparação de *wavelets* de Gabor, que rastreiam 18 pontos chave da face. Dados esses pontos, a expressão é reconhecida por meio de uma Rede Bayesiana acoplada que permite unificar a análise e a síntese das expressões para obtenção dos parâmetros de animação. São utilizadas as distribuições de probabilidade de seis expressões modelo. De acordo com os autores, a vantagem da abordagem com redes Bayesianas é garantir a robustez da animação, mesmo quando o rastreamento falhar. Uma das contribuições desse trabalho é que existe uma grande compressão dos dados para animação, de forma a possibilitar a animação facial de avatares com baixas taxas de transmissão de dados.

Tang e Huang [73] propuseram uma metodologia para obtenção de parâmetros do padrão MPEG-4 a partir de Tabelas de Animação Facial. Nessa abordagem, o rastreamento de pontos chave é feito pelo ajuste de um modelo deformável 3D ao rosto do usuário, utilizando fluxo ótico e comparação de modelos 2D (*template matching*) com correlação normalizada. A partir dos deslocamentos dos pontos chave, os autores podem identificar Unidades de Ação (ver Seção 3.1). A presença de determinado conjunto de Unidades de Ação pode significar uma expressão facial mais global, definidas em Tabelas de Animação Facial. Após a identificação da expressão, parâmetros MPEG-4 mais baixo nível pré-definidos são transmitidos ao sistema de animação. A Figura 2.4 ilustra o processo e mostra resultados obtidos pelos autores.

Cao e colaboradores [8] apresentaram um sistema para animação dirigida por performance em tempo real baseado em regressão 3D de forma. Nesse sistema, as posições tridimensionais de pontos de controle são estimadas por um regressor a partir das posições 2D obtidas em quadros de vídeo. A partir desses pontos tridimensionais, a pose da cabeça e as expressões são recuperadas através de ajuste a um modelo de *blendshapes* (ver Seção 3.4.2) específico do usuário. Para construção do modelo de face específico do usuário, uma fase inicial de configuração é requerida. Nessa fase, é solicitado ao usuário que gire sua cabeça em 15 ângulos diferentes. Após, é solicitado que o usuário realize 15 movimentos faciais pré-estabelecidos em três diferentes ângulos de rotação em torno do eixo y. Com o resultado do aprendizado do formato de faces 3D a partir de um banco de dados de faces digitalizadas, constrói-se o avatar do usuário. Um modelo de rastreamento já utilizado em trabalhos anteriores dos mesmos autores [83], baseado em mistura de Gaussianas, é utilizado. É possível realizar o mapeamento do rastreamento para avatares diferentes do usuário, conforme pode ser observado na Figura 2.5.

De acordo com os trabalhos apresentados nessa seção, há um grande avanço no campo de pesquisa em PDA, tendo sido obtidos resultados encorajadores. Especialmente em [8], são relatados

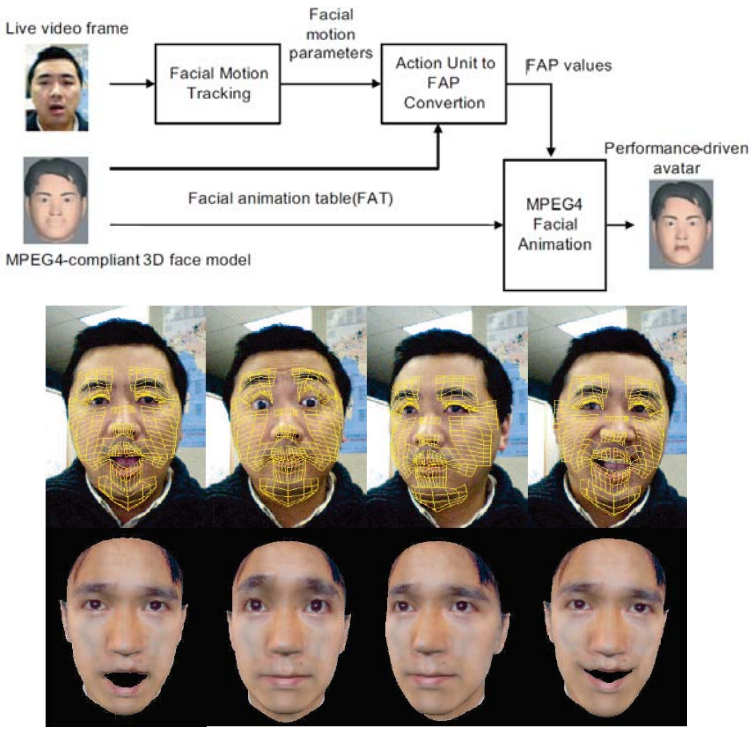


Figura 2.4: Esquema e resultados obtidos pela abordagem proposta por Tang e Huang [73]

resultados excelentes no aprendizado e mapeamento de expressões personalizadas a partir de sequências de imagem são relatados. Porém, não foi encontrada referência à animação facial de avatares dirigida pela performance de um usuário com expressões personalizadas de outro ator.

Conforme mencionado anteriormente, o reconhecimento automático de unidades de ação é essencial para a estratégia proposta para abordagem do problema de pesquisa desta tese. A próxima Seção apresenta alguns dos trabalhos nessa área.

2.2 Reconhecimento Automático de Unidades de Ação

Essa seção visa apresentar resultados obtidos na área de reconhecimento de expressões faciais. Dada a vasta extensão das pesquisas feitas nesse campo, essa revisão será limitada a modelos que visem reconhecer as unidades de ação aparentes na face, descritas no sistema de codificação de ações faciais proposto por Ekman [18] e apresentado na Seção 3.1. Assim, não serão apresentados trabalhos que tenham por objetivo classificar a expressão facial global (reconhecimento de alegria, raiva, asco, surpresa, medo e tristeza). A revisão também será restrita à reconhecimento de unidades de ação em imagens ou sequências de vídeo obtidas por câmeras monoculares. Assim sendo, não serão considerados trabalhos referentes ao reconhecimento de ações faciais em dados tridimensionais.

De forma geral, o processo de classificação de unidades de ação faciais em imagens ou sequências de vídeo pode ser dividido em três etapas:

- Extração de características relevantes, que podem ser o formato de componentes facias, presença de características transientes como rugas, abertura dos olhos, deslocamentos em relação

Figura 2.5: Resultados apresentados em [8]. Da esquerda para a direita: pontos rastreados no rosto do usuário, modelo 3D específico do usuário, mapeamento das expressões para um avatar semelhante a humanos e mapeamento de expressões para um avatar não semelhante a humanos.

à face neutra, características da textura, entre outros;

- Treinamento de classificadores: classificadores automáticos são treinados, geralmente por meio de aprendizado supervisionado, para associar às características de entrada, unidades de ação da face;

- Classificação: após a etapa de treinamento, os classificadores podem ser utilizados para identificar a presença de unidades de ação dados novos vetores de características obtidos em novas imagens de faces.

Yang e colegas [88] utilizaram características *Haar-like* [55] extraídas de sequências de imagens normalizadas para codificar deslocamentos da face neutra para cada uma das unidades de ação. Classificadores *Adaboost*[1] foram utilizados para discriminação de unidades de ação de acordo com o FACS.

Zhu e colegas também utilizaram classificadores *Adaboost* em [91]. Eles propuseram uma nova abordagem para combinação de classificadores em cascata com *bootstrapping*[2] bidirecional. As características utilizadas para classificação foram obtidas por modelos de aparência ativa (descritos na Seção 3.2) e descritores SIFT (do inglês *Scale Invariant Features Transform* [40]). Os autores compararam seu método com métodos alternativos usando máquinas de suporte vetorial treinadas com características provenientes da aplicação de *wavelets* de Gabor [35] e relataram resultados superiores.

[1] Do inglês *Adaptative Boosting* - técnica de combinação de classificadores fracos gerados subsequentemente de forma a favorecer exemplos erroneamente classificados anteriormente [4]

[2] O termo *bootstrapping* se refere a uma técnica para combinação de classificadores que tem por objetivo obter uma saída mais acurada do que o melhor dos classificadores individuais. Nessa técnica, os dados de entrada são reamostrados com reposição e são geradas classificações individuais para cada amostra. A escolha da classe final será definida por sistema de votos [4].

Hamm e colaboradores [24] treinaram 15 classificadores *Adaboost* independentes para detecção de unidades de ação. Como características de entrada, esses classificadores recebem características de textura e de forma. As características de textura são obtidas pela aplicação de *wavelets* de Gabor enquanto que as características de forma foram obtidas pela subtração das imagens de cada expressão em relação à face neutra. Os autores relataram taxas de sucesso que variam de 87% a 99,3% na classificação, dependendo da unidade de ação. A análise temporal das unidades de ação presentes em vídeos de faces de pacientes foi utilizada para auxiliar o diagnóstico de doenças neuropsiquiátricas.

Senechal e colegas utilizaram histogramas de padrões binários locais de Gabor para extração de características de imagens de faces alinhadas e reescaladas em [66]. Também utilizaram modelos de aparência ativos (AAM). Os classificadores utilizados foram Máquinas de Suporte Vetorial (SVM - do inglês *Support Vector Machines*) com múltiplos núcleos. Para treinamento das SVMs, os autores utilizaram, entre outros, o banco de dados Bosphorus, descrito na Seção 3.5. Na comparação de resultados das características de entrada dos classificadores, os autores reportaram um desempenho superior dos histogramas de padrões binários locais de Gabor em relação ao AAM.

Chu et al. [10] propuseram uma técnica para personalizar classificadores de unidades de ação. Os autores utilizaram uma técnica transdutiva chamada Máquina de Transferência Seletiva. A principal ideia por trás dessa técnica é atribuir um peso maior aos exemplos do banco de dados de treinamento que são mais próximos da amostra de teste. Máquinas de Suporte Vetorial foram utilizadas como classificadores.

A fim de mostrar a diversidade de técnicas e a variedade de trabalhos na área, alguns outros exemplos serão citados brevemente na Tabela 2.1, informando as técnicas de extração de características de entrada e os classificadores utilizados.

Tabela 2.1: Métodos de extração de características de entrada e classificadores utilizados em alguns trabalhos de reconhecimento de unidades de ação facial.

Autores	**Extração de Características**	**Técnica de Classificação**
Pantic et al. [54]	Segmentação por cor	Raciocínio Baseado em Regras
Li et al. [37]	Pontos obtidos por ASM	Redes Bayesianas dinâmicas e AdaBoost
Chuang e Shih [11]	Análise de Componentes Independentes	SVM
Mahoor et al. [44]	Filtros de Gabor, minimização de norma L_1	Maximização de ordenamento
Wu et al. [85]	Filtros de Gabor, padrões binários locais	Uma e Duas camadas de SVMs
Gonzalez et al. [22]	Filtros de Gabor	SVM e *AdaBoost*

A próxima seção descreve brevemente alguns trabalhos que utilizaram redes neurais artificiais para classificação de unidades de ação. Esses trabalhos estão em uma seção separada, pois essa técnica foi utilizada no modelo desenvolvido nesta tese.

2.2.1 Utilização de Redes Neurais Artificiais para Classificação de Unidades de Ação

Os trabalhos descritos a seguir utilizaram redes neurais artificiais para classificação de unidades de ação.

Tian e colegas treinaram redes neurais artificiais para reconhecimento de unidades de ação a partir de características permanentes (olhos, boca, sobrancelha) e transientes (buracos e rugas) detectadas na face em [75]. A classificação é feita com base no formato dos componentes faciais e presença ou ausência de características transientes. Para detecção do formato da boca, os autores utilizaram mistura de Gaussianas para representar a distribuição de cor dos lábios. Quanto aos olhos, os autores basearam-se na porcentagem aparente da íris, representada por um círculo. A presença de rugas nas regiões nasolabial e subocular é definida pela aplicação de um detector de bordas. Para classificação das características obtidas, os autores treinaram redes neurais artificiais diferentes para as unidades de ação da parte inferior da face e para a parte superior. Foi obtido sucesso em 88,5% das unidades de ação relacionadas com a parte inferior da face e 93% para as unidades de ação relacionadas com a parte superior da face.

Kuilenburg et al. [80] utilizaram redes neurais do tipo *feedforward* para reconhecimento de unidades de ação da face, utilizando como características de entrada resultados de detecção por AAM (Seção 3.2). O sistema proposto também é capaz de reconhecimento das seis emoções básicas. O reconhecimento de unidades de ação teve um percentual de acerto de 86% quando aplicado na base de dados Cohn-Kanade [29].

Seyedarabi aplicaram redes neurais probabilísticas para classificação de nove unidades de ação dos lábios em [67]. Eles desenvolveram um sistema para rastreamento dos lábios em imagens baseado em modelos de contorno ativos. Resultados experimentais apresentaram um desempenho médio de classificação de 85,98% no banco de dados Cohn-Kanade.

Kotsia e colegas utilizaram redes neurais artificiais para combinar a classificação de forma e textura obtidas por máquinas de suporte vetorial em [32]. Para extração de características de forma, foi utilizada a máscara Candide [2]. Já para extração de informações de textura foi utilizada a fatoração em matrizes de discriminantes não negativos. Para a obtenção de ambas as características foi realizada a diferença entre cada expressão e a face neutra. O resultado da classificação por SVM tanto para as informações de variação de textura quanto para as informações de variação são submetidos a um subsistema de fusão para a classificação final. Os autores testaram vários métodos e concluíram que uma rede neural com função de base radial atingiu bons resultados. Os autores obtiveram 92,1% de sucesso na classificação de 17 unidades de ação faciais quando aplicado ao banco de dados Cohn-Kanade [29].

2.3 Estilo de Movimento

Nos últimos anos, dispositivos de aquisição de informação sobre o movimento de atores (*motion capture*) têm permitido a pesquisadores obter dados detalhados sobre o movimento humano. Com essa disponibilidade de dados, o interesse pela síntese de movimento para animação de personagens

baseada em aprendizado tem se intensificado. Além de aprender parâmetros gerais do movimento que identifique um conteúdo (andar, pular, correr, etc) pode-se pensar também em aprender as particularidades dos movimentos de um ator específico, um sentimento expressado pela forma de executar o movimento, ou mudanças que dependem do gênero do ator. Segundo Torresani e colegas [52], tais características são denominadas estilo. Essa seção descreve brevemente trabalhos cujo objetivo é o aprendizado do estilo de movimento a partir de dados de captura de movimento, para síntese de animação de personagens virtuais.

Pan e Torresani [52] descreveram uma técnica de aprendizado não supervisionado para modelar estilos de locomoção humanos e variações do mesmo movimento realizadas por diferentes sujeitos. Segundo os autores, a modelagem de estilos de movimento requer a identificação da estrutura comum do movimento e a detecção de características específicas de estilo. O modelo proposto neste trabalho é restrito a movimentos cíclicos como caminhar, correr ou nadar. O estilo é modelado como uma variável aleatória que deriva de um "pai" comum representando a estrutura compartilhada por todos os movimentos com mesmo conteúdo. Com as trajetórias cíclicas desses dados alinhadas, os autores descobriram variáveis de estilo de forma não supervisionada.

Liu et al. [39] propuseram um modelo para síntese de movimento humano com vários estilos, a partir de dados de captura de movimento. Nesse modelo, as diferenças entre os estilos de dada ação são expressas como um subespaço dos dados capturados, que os autores denominaram subespaço de estilo. Eles utilizaram a técnica de análise de subespaços independente de características [33] para encontrar os subespaços de estilo. Com esses subespaços, um grupo de modos de edição foi desenvolvido para que animadores possam não somente modificar o estilo de um movimento original capturado mas também transferir e mesclar estilos.

Etemad e Arya [20] propuseram um modelo para reconhecimento de ações humanas baseados em *Hidden Markov Models - HMM [4]*. Uma vez reconhecida a atividade principal (conteúdo), os autores mostraram como variar o estilo dessa atividade de forma paramétrica. O objetivo é converter, por exemplo, a caminhada de um ator masculino cujo movimento tenha sido capturado, na caminhada de um personagem virtual feminino. Para atingir esses objetivos, os autores utilizaram duas técnicas: interpolação e *warping* temporal não linear.

Taylor e Hinton utilizaram máquinas de Boltzmann restritas condicionais, que é um modelo para aprendizado de transições entre séries temporais, para aprendizado e síntese de estilos de movimentos em [74]. Os autores salientam que um dos objetivos durante o aprendizado dos parâmetros é separar o estilo (por exemplo, tristeza) do conteúdo do movimento (por exemplo, caminhar do ponto A ao ponto B). O modelo proposto pelos autores é capaz de representar diversos estilos de movimento por meio de um único conjunto de parâmetros e tem as habilidades de combinar diferentes estilos de movimento e permitir uma transição suave entre eles. Para aprendizado, os autores utilizaram um banco de dados de captura de movimentos contendo 10 estilos anotados *a priori*.

Min e colaboradores [47] apresentaram um modelo de geração de movimentos humanos para síntese, reaplicação e edição de estilos de movimento personalizados. Primeiramente, os autores gravaram um banco de dados de múltiplos atores realizando uma variedade de estilos para ações

particulares. Eles aplicaram técnicas de análise multilinear para construção desse modelo, com parâmetros de controle para variação de estilo e identidade. Como resultados, esses parâmetros podem ser usados para sintetizar novos movimentos, transferir o estilo de movimento de um ator para outro e editar o movimento. Imagens com resultados atingidos pelos autores são mostradas na Figura 2.6. A imagem da esquerda mostra um estilo de caminhada usado para estimar parâmetros de identidade; a imagem central mostra o movimento de caminhar sorrateiro do mesmo ator da imagem da esquerda. A imagem da direita mostra a caminhada sorrateira sintetizada pela combinação da assinatura do movimento do ator estimada pelo modelo e do estilo sorrateiro.

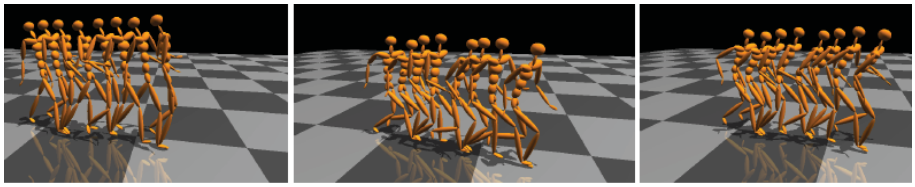


Figura 2.6: Resultados obtidos por Min e colegas [47].

Chiu e Marsella [9] aperfeiçoaram o método descrito em [74]. Eles propuseram um método para separação entre o conteúdo de movimento e seu estilo e interpolação de estilos de movimento. Especificamente, eles estenderam a técnica de máquinas de Boltzmann restritas fatoradas, tornando-as hierárquicas. A hierarquia consiste em utilizar uma camada oculta para interpolação entre estilos. O resultado da interpolação através da camada oculta é um vetor de movimento que pode ser diferente de todos os exemplos utilizados em treinamento. Como exemplo de aplicação os autores apresentaram a interpolação entre dois estilos de caminhada: marcha e caminhada rápida.

No contexto de estilo de movimento para animação facial especificamente, pode-se citar o trabalho de Lee e Samaras [34], que apresentaram uma técnica para modelar expressões faciais em múltiplas pessoas com diferentes expressões e sintetizar novas faces com expressões sutis estilizadas usando modelos geradores. Os autores modelaram as expressões faciais como mapeamento num espaço não linear. Características de diferentes tipos de expressão e variâncias entre diferentes pessoas são decompostas. Utilizando-se de rastreamento em alta resolução de dados tridimensionais densamente amostrados, o modelo pode controlar características sutis das expressões de diferentes pessoas. Resultados da síntese de sorrisos com várias sutilezas distintas podem ser observados na Figura 2.7.

2.4 Contextualização do Trabalho no Estado da Arte

Esta tese tem por objetivo o aprendizado do estilo de movimento facial para posterior utilização em aplicações de Animação Dirigida por Performance de usuários, com aquisição de dados por câmeras monoculares. Alguns trabalhos relacionados à PDA foram descritos na Seção 2.1. O levantamento bibliográfico feito para essa seção não revelou trabalhos com propósitos idênticos aos especificados nos objetivos desta tese. Pôde-se perceber a preocupação dos autores da área em que o movimento do avatar fosse o mais semelhante possível ao movimento do usuário. No entanto

Figura 2.7: Resultados obtidos por Lee e Samaras [34].

esses trabalhos não apresentaram a figura de um ator cujo estilo de movimento se queira utilizar. Portanto, verifica-se que há um campo de pesquisa em aberto no qual essa tese pretende contribuir.

Na abordagem proposta, é essencial o reconhecimento de unidades de ação tanto nas expressões dos usuários quanto nas expressões dos atores. Por isso, alguns trabalhos relacionados a esse campo de pesquisa foram apresentados na Seção 2.2. Foi escolhida, nesta tese, uma abordagem baseada em análises de componentes principais e redes neurais artificiais. Embora outras técnicas tenham também sido utilizadas nos últimos anos, foi escolhida a utilização de RNAs dado o conhecimento prévio dessa técnica por parte da autora da tese e dados os bons resultados reportados pelos autores citados na Seção 2.2.1. A investigação de outras técnicas demandaria tempo de aprendizagem e implementação. Porém, trabalhos futuros devem ser dedicados a essa investigação, especialmente no que diz respeito à utilização de SVMs com *AdaBoost*.

A Seção 2.3 mostra que há um campo ativo de pesquisa na área de aprendizado do estilo de movimento para síntese de animação e transferência de movimento entre diferentes modelos geométricos. Contudo, notou-se que o aprendizado do estilo de movimento facial particularizado somente tem sido abordado nos últimos anos, com apenas um trabalhos encontrado. Assim sendo, esta tese pretende contribuir também nesse campo de pesquisa.

O próximo capítulo apresentará alguns conceitos fundamentais necessários para a compreensão do modelo que está sendo proposto.

3. CONCEITOS FUNDAMENTAIS

Esse capítulo descreve técnicas e conceitos que foram utilizados no desenvolvimento desse trabalho. A próxima seção apresenta o *Facial Action Code System*, sistema proposto por Paul Eckman para classificação de unidades de ação na face, que se constituem de movimentos visíveis no rosto. Esse sistema será utilizado para classificação das expressões faciais e descrição do movimento do rosto tanto dos atores como dos usuários.

Como o modelo proposto depende do rastreamento em tempo real de pontos característicos da face tanto para o aprendizado do estilo de movimento facial do ator (Persona) quanto para a utilização desse estilo para a animação do avatar guiada pelo usuário, a Seção 3.2 revisa algumas técnicas da literatura usadas para esse fim e apresenta a ferramenta *Live Driver*, que será utilizada nesse trabalho. A Seção 3.3 descreve duas técnicas amplamente utilizadas na área de reconhecimento automático de padrões que são utilizadas para classificação das ações realizadas pelo ator ou usuário: Análise de Componentes Principais e Redes Neurais Artificiais. Já a Seção 3.4 mostra técnicas e parâmetros que podem ser utilizados para animar o avatar. A Seção 3.5 apresenta o banco de dados Bosphorus [64], o qual é constituído de imagens e nuvens de pontos anotadas da face de 105 pessoas, cada uma em diferentes expressões. A Seção 3.6 descreve o algoritmo para alinhamento de nuvens de pontos chamado Análise de Procrustes. Finalmente, a Seção3.7 apresenta uma medida de similaridade para comparação entre conjuntos de pontos.

3.1 Sistema de Codificação de Ações Faciais

O estudo de emoções expressas através da face existe desde a época de Charles Darwin que publicou análises de expressões faciais em seu livro intitulado *The Expression of the Emotions in Man and Animals* [15]. Nesse livro ricamente ilustrado, Darwin publica observações sobre o fato de que povos geograficamente isolados apresentam expressões semelhantes para as mesmas emoções. Embora alguns autores tenham encontrado indícios para contestar essa observação (por exemplo, Russel em [62]), Paul Ekman [18] encontrou evidências de que algumas expressões são universais e, portanto, devem ter uma origem fisiológica. Essas emoções são alegria, medo, asco, surpresa, tristeza e raiva.

Ekman catalogou os movimentos faciais fisiologicamente possíveis e perceptíveis a um observador, denominados "Unidades de Ação" (*Action Units - AU*). Em outras palavras, as AUs representam os movimentos musculares da face que são visíveis. Essa categorização encontra-se presente no "Sistema de Codificação de Ações Faciais" [18] (*Facial Action Coding System - FACS*). São, ao todo, 44 unidades de ação. A Figura 3.1 mostra a representação de grupos musculares próximos ao olho direito e as respectivas Unidades de Ação desses músculos. Por exemplo, nessa figura, a AU_1 consiste em erguer o canto interno da sobrancelha, enquanto a AU_4 consiste no movimento de unir as sobrancelhas.

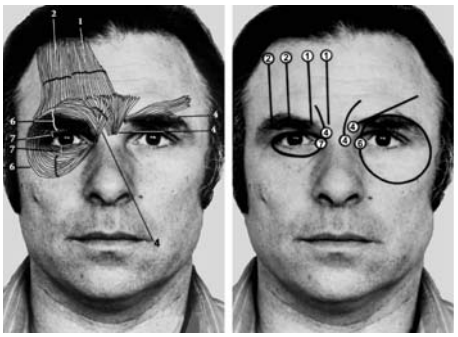


Figura 3.1: À esquerda, observam-se grupos musculares ao redor do olho direito. À direita, suas respectivas Unidades de Ação (AUs) [18].

Essas unidades de ação podem ser combinadas de forma mostrar expressões complexas. Algumas emoções ditas universalmente reconhecidas, são caracterizadas pela presença específica unidades de ação combinadas. Conforme mencionado anteriormente, essas emoções são alegria, tristeza, asco, surpresa, medo e raiva. A emoção alegria, por exemplo, é representada pela combinação da unidade de ação 6 (elevação da bochecha e contração do músculo orbicular do olho) e a unidade de ação 12 (tração do canto da boca lateralmente e para cima). Essas unidades de ação podem ser visualizadas nas Figuras 3.21 e 3.22 e suas combinações para expressar as seis emoções básicas universais podem ser vistas na Figura 3.20. Tais imagens foram retiradas do Banco de dados de faces Bosphorus, descrito na Seção 3.5.

Através do FACS, Ekman criou uma metodologia para reconhecimento e quantificação das AUs. A partir dessas quantificações, ele categorizou as expressões de acordo com as combinações e intensidades das Unidades de Ação presentes.

Esse sistema serviu como fundamentação para diversos trabalhos de reconhecimento de expressões faciais por Visão Computacional citados ao longo do presente trabalho. Além disso, o padrão de compactação de dados MPEG-4 utilizou-o como base para representação da animação de faces de modelos 3D. A Seção 3.4.1 apresenta uma sucinta descrição desse padrão.

3.2 Rastreadores de componentes faciais

Os processos de animação da face de um avatar e classificação de expressões faciais requerem, necessariamente, uma boa estratégia para detecção e rastreamento de pontos significativos como os cantos dos olhos, pálpebras, íris, sobrancelhas, narinas e pontos no contorno dos lábios, os quais serão chamados de pontos característicos. Tais ferramentas devem processar sequências de imagens em tempo real, com boa robustez à variação de iluminação e de resolução da face. Essa é uma tarefa desafiadora dado o grande número de graus de liberdade da face humana, a possibilidade de mudança na iluminação, rapidez dos movimentos e eventuais oclusões devidas a gestos, cabelos ou rotação da cabeça [86]. Esses parâmetros são mais facilmente adquiridos com os equipamentos sofisticados de *motion capture*, quando os atores utilizam marcadores na face, mas são difíceis de detectar e rastrear em imagens monoculares. Essas dificuldades podem ainda ser maiores se o

dispositivo com a qual foi adquirido o vídeo tiver baixa resolução, como é o caso de *webcams* ou de câmeras com lentes de grande ângulo de abertura. Embora essa seja uma tarefa desafiadora, vários pesquisadores têm se dedicado a prover o rastreamento estável desses pontos nos últimos anos. Essa seção apresenta alguns dos trabalhos relevantes nessa área.

Os Modelos de Forma Ativos (*Active Shape Models* - ASM) foram propostos por Cootes e colegas para análise e rastreamento de objetos deformáveis [13]. Essa é uma abordagem baseada em treinamento, na qual deve ser construído um modelo de distribuição de pontos que descreve o contorno do objeto que se deseja rastrear, detectar ou segmentar. Esse modelo de distribuição de pontos deve ser anotado em um conjunto de imagens contendo o objeto de interesse em vários formatos. No caso da face, deve-se construir o modelo de distribuição de pontos considerando os pontos característicos do rosto em diversas expressões distintas. Na fase de treinamento, os modelos de distribuição de pontos anotados são escalados e alinhados para que seja aprendida a sua forma média ($\bar{\mathbf{s}}$) e para que sejam obtidos os seus modos de variação possíveis, utilizando Análise de Componentes Principais. Esses modos de variação são dados pelos t auto-vetores $\mathbf{P}$ da matriz de covariância dos dados de treinamento . Assim, uma dada forma $\mathbf{s}$ pode ser aproximada por:

$$\mathbf{s} \approx \bar{\mathbf{s}} + \mathbf{P}\,\mathbf{b} \tag{3.1}$$

onde $\mathbf{b}$ é um vetor t-dimensional dado por $\mathbf{b} = \mathbf{P}^T(\mathbf{s} - \bar{s})$. Variando-se os elementos de $\mathbf{b}$, pode-se variar a forma obtida $\mathbf{s}$. A variância do i-ésimo parâmetro de $\mathbf{b}$ ($\mathbf{b}_i$) considerando os dados de treinamento é dada por λ_i. Aplicando-se os limites de $\pm 3\sqrt{\lambda_i}$ ao parâmetro $\mathbf{b}_i$ assegura-se que a forma gerada é similar àquelas do conjunto original de dados de treinamento.

Uma das abordagens mais utilizadas atualmente para rastreamento dos pontos significativos foi a extensão dos ASMs proposta por Cootes et al. em [12]. Nesse trabalho, foi desenvolvido um modelo estatístico para descrever os modos de variação da forma e da aparência em escala de cinza de objetos de interesse, os quais são conhecido como Modelos de Aparência Ativos (*Active Appearance Models – AAM*). Com isso, informações de textura são utilizadas juntamente com os ASMs para o rastreamento de objetos deformáveis. Nessa técnica, é também requerido um conjunto de imagens anotadas, nas quais pontos-chave da face devem ser marcados manualmente. Dado esse conjunto, na fase de aprendizado é construído um conjunto de modos de variação ortogonais, dados pelos auto-vetores da matriz de covariância, obtidos por Análise de Componentes Principais. A técnica apresenta bons resultados quando o objeto rastreado tem formato e aparência semelhantes a objetos no conjunto de treinamento, mas é sensível a variações de iluminação e a condições ausentes no conjunto de dados de treinamento. Para contornar esses aspectos negativos, é necessário um grande banco de dados para aprendizagem, que apresente grande variabilidade de sujeitos, expressões e condições de iluminação. Entretanto, a marcação manual que deve ser feita para a fase de aprendizado é exaustiva e, para um melhor resultado no caso específico de faces humanas, deve ser feita para cada indivíduo cujo movimento se deseja mapear. Dentre os trabalhos que lançaram mão dos AAMs para rastreamento de pontos chave da face, podemos citar [3], [7], [23], [38], [77] e [89].

Os autores ainda adaptaram os AAMs para rastreamento de componentes faciais em dispositivos móveis em [79].

Uma outra abordagem muito utilizada para rastreamento de faces em imagens é a utilização de Modelos Deformáveis 3D (3D Morphable Models - 3DMM). Esses modelos consistem em máscaras de polígonos que representam a face humana e que podem ser deformadas de acordo com as particularidades do rosto do usuário ou com as expressões faciais apresentadas. De forma geral, esse processo passa por uma etapa de inicialização em que deve ser feita a correspondência entre os vértices do modelo e pontos chave do rosto do usuário na imagem. Essa inicialização requer que os pontos chave da face na imagem sejam conhecidos e corretamente associados aos vértices do modelo correspondentes. Em seguida, essa máscara é deformada de forma a se ajustar à face do usuário. Após, são usados algoritmos de rastreamento que buscam manter esse ajuste da melhor forma possível, tanto deformando o modelo quanto rotacionando-o e transladando-o. Uma grande vantagem dessa técnica é que o rastreamento é eficiente mesmo com oclusões parciais da face do usuário devido à rotação, pois se lança mão da projeção do modelo do espaço 3D para o espaço 2D e conhece-se quais pontos devem estar visíveis na imagem, mesmo com a cabeça rotacionada. DeCarlo e Metaxas [16] utilizaram a análise de fluxo ótico da imagem para rastreamento de pontos chave e seus resultados podem ser vistos na Figura 3.2.

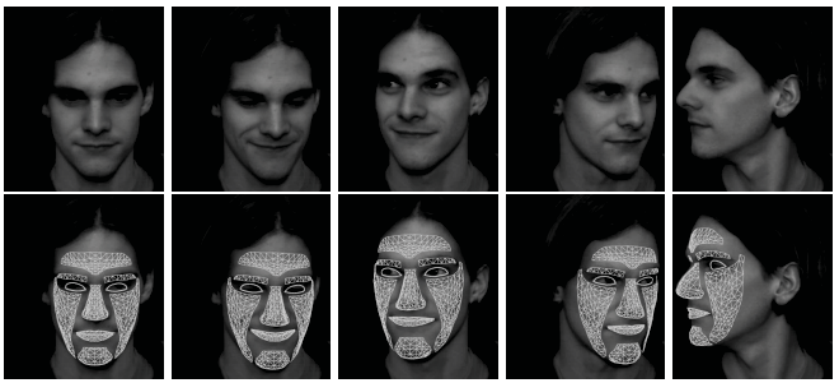


Figura 3.2: Resultados obtidos em [16]. Pode-se observar a correta correspondência do modelo 3D ao rosto do usuário mesmo com oclusão parcial devida à rotação.

A título de exemplificação, seguem mais alguns trabalhos que utilizaram ajuste de modelos 3D para rastreamento de faces e pontos chave faciais. Schramm [65] utilizou o modelo de face Candide [1] e atributos de cor da imagem para o rastreamento rígido da pose da cabeça utilizando o algoritmo Lucas–Kanade–Tomasi (LKT) [41]. Vogler e Goldenstein [81] utilizaram rastreamento tanto rígido quanto não rígido com modelos deformáveis para prover leitura labial e de expressões para interpretação de sinais para deficientes auditivos. Mpiperis et al. [49] propuseram um Modelo Deformável Bilinear para reconhecimento de expressões faciais.

Cristianacce et al. descreveram um método iterativo para encontrar pontos chave em imagens estáticas e em vídeo [14]. O algoritmo desenvolvido realiza correspondência de modelos 2D (*template matching*) de partes da face otimizado por restrições impostas por um modelo da forma da face. Os modelos das partes da face são atualizados em tempo de execução para garantir robustez ao longo

do tempo. Os testes feitos pelos autores mostram bons resultados no rastreamento de 20 pontos chave. Entretanto, esses pontos não incluem as pálpebras superiores e inferiores e apenas quatro pontos descrevem o movimento da boca.

Sohail e Bhattacharya desenvolveram um modelo de face baseado em estatísticas antropométricas para detecção dos 18 pontos-chave mais importantes [70]. Nesse modelo, a distância entre as duas pupilas obtida através de técnicas de detecção e classificação de objetos proposta por Fasel e colegas [21] serve como principal parâmetro de medida para a localização do centro dos outros componentes faciais. Os autores utilizaram operações morfológicas, binarização e detecção de bordas para obtenção de características que foram submetidas a um conjunto de regras geométricas e baseadas na intensidade dos pixels para localização de cada componente facial que os autores consideraram relevante. Por exemplo, para detecção dos olhos, binariza-se a região próxima à localização da pupila, determina-se componentes conexos utilizando um elemento estruturante de 4 pixels e, para o olho direito, determina-se que o canto direito do componente conexo é o canto direito do olho.

Ong et al. propuseram uma abordagem para rastreamento em tempo real de pontos chave da face em [51]. Nessa abordagem, restrições impostas por modelos de forma ou modelos temporais para a dinâmica dos movimentos faciais são desnecessárias. De acordo com os autores, o rastreamento é feito via Preditores Lineares (*Linear Predictors*) que consideram informações dos próprios pixels da imagem para rastreamento do deslocamento das regiões onde se encontram os pontos chave. Resultados do rastreamento usando essa abordagem podem ser vistos na Figura 3.3. Chase e colegas utilizaram árvores de regressão para aprimorar os resultados, construindo preditores não lineares [69].

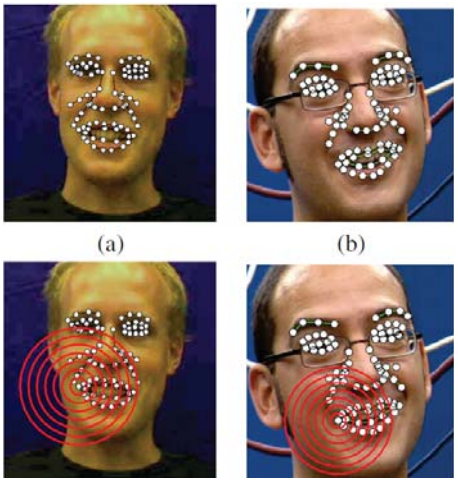


Figura 3.3: Resultados de rastreamento de pontos chave obtidos em [51].

Os métodos descritos nessa seção apresentaram bons resultados, considerando que utilizam como informações de entrada apenas imagens monoculares. Essa listagem de metodologias não é exaustiva e visa mostrar diferentes abordagens para rastreamento de componentes faciais. Durante o desenvolvimento desse trabalho, porém, a empresa *Image Metrics* lançou o *kit* de desenvolvimento de software *Live Driver*. Conforme informações apresentadas na próxima seção, essa é uma ferramenta

que provê a detecção e rastreamento de pontos característicos da face em tempo real, com boa acurácea mesmo em condições não ideais de iluminação, oclusão parcial ou rotação da face. Por isso, decidiu-se utilizar essa ferramenta para prover os dados que servirão de entrada para o modelo proposto no Capítulo 4. Os demais métodos de rastreamento apresentados nessa seção mostramn que o modelo proposto não é dependente do *Live Driver* e que existem alternativas não comerciais para provimento dos dados de entrada.

3.2.1 A Ferramenta de Desenvolvimento de *Software Live Driver*

O *Kit* de Desenvolvimento de *Software Live Driver* da *Image Metrics*[1] é uma ferramenta de análise da face, rastreamento de pontos característicos e fornecedor de parâmetros de animação, disponibilizada como uma biblioteca em C++ para computadores pessoais. O rastreamento dos pontos característicos das faces mostrados na Figura 3.4 ocorre em tempo real (30 quadros por segundo), em condições variáveis de resolução da imagem e iluminação. É uma tecnologia proprietária, com poucas informações disponíveis sobre os modelos teóricos utilizados.

A Figura 3.4 mostra imagens de faces em diferentes resoluções, ângulos da cabeça e condições de iluminação, com pontos característicos detectados e rastreados pelo *Live Driver*. Na imagem da esquerda, a resolução da face é de aproximadamente 340×360 pixels enquanto na imagem central a resolução fica em torno de 100×120 pixels. A imagem da direita da mesma figura mostra bons resultados de rastreamento mesmo com rotação da face e sobreposição de outra imagem ao fundo.

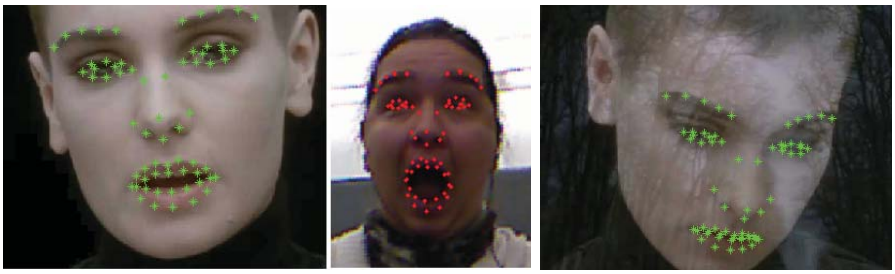


Figura 3.4: Pontos característicos rastreados pela ferramenta *Live Driver*.

Além dos 64 pontos característicos mostrados na Figura 3.4, o *Live Driver* fornece a posição da cabeça, uma estimativa dos ângulos de rotação da cabeça em termos de ângulos de Euler (α, β e γ, conforme Figura 4.4) e um conjunto de parâmetros de controle de animação, de forma semelhante ao padrão MPEG-4 explicado na Seção 3.4. Para os propósitos desse trabalho, serão utilizados apenas as coordenadas dos pontos característicos e os ângulos de rotação (α, β e γ).

Segundo a empresa desenvolvedora, o *Live Driver* foi otimizado para se adequar a uma grande variedade de aparências, incluindo diversidade de raça e idade e presença de barbas, bigodes ou óculos. A título de comparação com o desempenho de outra ferramenta comercial, a Figura 3.5 mostra os resultados do *Live Driver* comparados com os resultados obtidos pelo kit de desenvolvimento de

[1] `http://www.image-metrics.com/livedriver/overview/`, consultado em Julho de 2014

software do dispositivo Kinect[2] da *Microsoft*. Nesse último *kit*, os dados de rastreamento são também obtidos em tempo real. Porém, o *kit* da *Microsoft* utiliza informação do mapa de profundidades fornecido pelo *Kinect* conjuntamente com os dados de imagem. Note que, mesmo utilizando mapas de profundidade, o resultado de rastreamento do *kit* do Kinect mostrado nas imagens à esquerda da Figura 3.5 é inferior ao resultado do *Live Driver*, mostrado à direita. As imagens de entrada de ambos os *kits* são rigorosamente as mesmas, pois foi fornecida a informação capturada pela câmera do *Kinect* ao *Live Driver* (mostradas no centro na figura). Nas imagens da linha superior, o *kit* de rastreamento do *Kinect* falha na detecção do lábio inferior; nas imagens da linha superior, ele detecta os olhos como se estivessem completamente abertos. O *Live Driver* obteve resultados satisfatórios em ambos os casos.

Figura 3.5: À esquerda, pontos característicos obtidos pelo rastreador de faces do Kinect. No centro, as imagem originais. À direita, resultado obtido utilizando-se o *Live Driver*.

A ferramenta *Live Driver* pode ser integrada ao kit de desenvolvimento do Kinect e seus resultados podem ser mapeados para o mapa de profundidades RGBD (do inglês *Red, Green, Blue and Depth*) gerado por esse dispositivo. Resultados dessa integração podem ser vistos na Figura 3.6. Igualmente, o *kit Live Driver* pode ser utilizado em dispositivos móveis com os sistemas operacionais IOS ou Android. Assim sendo, o método desenvolvido nesse trabalho pode ser utilizado também usando dados capturados pelas câmeras desses dispositivos. Li e colegas utilizaram essa ferramenta aplicada em dados adquiridos pelo *Kinect* em seu sistema de animação dirigida por performance livre de calibração, descrito em [36].

Uma característica relevante do *Live Driver* é que o desempenho do rastreamento tende a melhorar com o tempo. Provavelmente é utilizada coerência temporal entre os quadros de vídeo para estabilizar e otimizar os resultados. Em geral, verificou-se que são necessários 20 quadros iniciais para que o rastreamento comece a ser estável e com uma taxa de sucesso maior.

Dadas as características apresentadas, decidiu-se que os dados obtidos pelo *Live Driver* serão utilizados como dados de entrada para o aprendizado da Persona de atores e para direção da animação do avatar por usuários quaisquer. Dessa forma, apenas informações de vídeo serão requeridas. Isso

[2] `http://www.microsoft.com/en-us/kinectforwindows/`, consultado em Julho de 2014

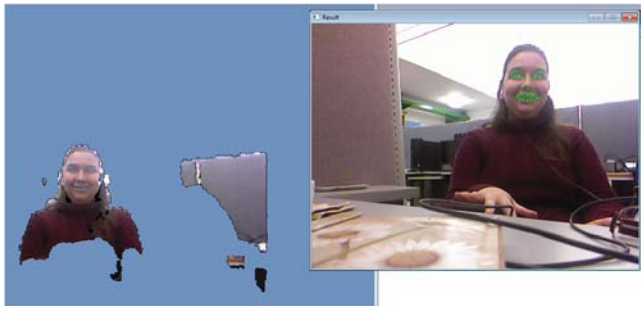


Figura 3.6: *Live Driver* realizando o rastreamento dos pontos nos dados capturados pela câmera do Kinect (à direita). Esses dados podem ser mapeados para a nuvem de pontos RGBD (à esquerda).

possibilita o aprendizado do estilo de movimento de atores sem necessidade de marcadores em suas faces para o rastreamento de pontos característicos e dispensa o uso de equipamentos mais sofisticados de captura de movimento. Assim, pode-se aprender o estilo de movimento de atores ausentes ou até mesmo falecidos, conforme explicado na Seção 1.2, apenas a partir de vídeos de suas expressões faciais.

De acordo com o que foi brevemente mencionado anteriormente, a Persona dependerá da classificação automática das ações rastreadas nas imagens. A próxima seção se destina a fornecer informações sobre as técnicas empregadas para esse fim.

3.3 Reconhecimento de Padrões

Para a construção da Persona de um ator A e sua posterior utilização para animação de um avatar de A dirigida pela performance de um usuário U, é importante que as unidades de ação propostas por Ekman (Seção 3.1) sejam corretamente identificadas. Para isso, os dados rastreados pelo *Live Driver* devem ser classificados automaticamente de acordo com as AUs presentes na performance do ator A ou do usuário U. Essa seção apresenta as técnicas de reconhecimento de padrões utilizadas nesse trabalho para tal finalidade, que consistirão de duas etapas: redução de dimensionalidade dos dados por Análise de Componentes Principais e classificação por Redes Neurais Artificiais.

3.3.1 Análise de Componentes Principais

Segundo Bishop [4], muitos conjuntos de dados têm a propriedade de que os pontos correspondentes aos dados num espaço de n dimensões ficam concentrados em um *manifold*[3] de dimensionalidade muito menor. Encontrar esse *manifold* pode reduzir, nesses casos, a complexidade dos dados e tornar mais fácil a extração de conhecimento a partir deles. A Análise de Componentes Principais (*Principal Component Analysis-PCA*) é uma técnica amplamente usada para esse fim em aplicações como redução de dimensionalidade, compressão de dados, extração de características relevantes e visualização de informação.

[3]Espaço topológico que se assemelha ao espaço Euclidiano na região próxima a cada ponto. A superfície de uma esfera é um exemplo de *manifold* bidimensional [26].

A PCA pode ser definida como a projeção ortogonal dos dados em um espaço linear de dimensionalidade menor, conhecido como **subespaço principal**, tal que a variância dos dados projetados é maximizada. Cada i-ésimo vetor da base desse novo subespaço é chamado Componente Principal (PC) e denotado por $\mathbf{u_i}$. É importante salientar que os PCs são vetores unitários e ortogonais entre si, constituindo assim uma base do subespaço principal. Dessa forma, cada vetor de dados de entrada $\mathbf{x_n}$ tem uma representação no subespaço principal $\tilde{\mathbf{x}}_n$ por uma combinação linear dos PCs, na forma:

$$\tilde{x}_n = \sum_{i=1}^{d} a_i.\mathbf{u}_i, \tag{3.2}$$

onde d é a dimensionalidade do subespaço principal. Se d for igual ao número de variáveis de entrada D, $\tilde{\mathbf{x}}_n$ é aproximadamente equivalente a $\mathbf{x}_n$, porém num novo sistema de coordenadas.

Considere os pontos vermelhos do gráfico da Figura 3.7 como dados $\mathbf{x}_n$ referentes a duas variáveis x_1 e x_2. O subespaço principal de dimensão $d = 1$ é, nesse caso, representado pela linha roxa e o seu PC é denotado por $\mathbf{u}_1$. Esse subespaço é definido de forma a maximizar a variância dos pontos nele projetados (em verde).

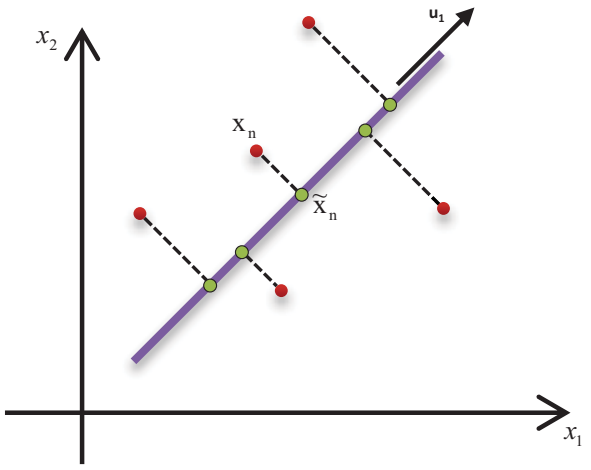


Figura 3.7: Projeção de dados de entrada $\mathbf{x_n}$ no componente principal $\mathbf{u_1}$.

Considere agora um conjunto de dados $\mathbf{x}_n$, onde $n = 1, 2, ..., N$ e $\mathbf{x}_n$ é uma variável Euclidiana com dimensionalidade D. Conforme explicado acima, o objetivo é projetar os dados em um espaço com dimensionalidade $d < D$ enquanto se maximize a variância dos dados projetados. Em um primeiro momento, seja $d = 1$. Pode-se definir a direção deste espaço usando um vetor unitário $\mathbf{u_1}$. Cada ponto $\mathbf{x}_n$ é então projetado em um valor escalar $\mathbf{u_1}^T\mathbf{x}_n$. A variância dos dados projetados é dada por:

$$\frac{1}{N}\sum_{n=1}^{N}(\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}})^2 = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1, \tag{3.3}$$

onde $\mathbf{S}$ é a matriz de covariância dos dados definida por:

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T, \tag{3.4}$$

com $\bar{\mathbf{x}}$ sendo a média amostral dos dados.

Deve-se agora maximizar a variância projetada $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$ (Equação 3.3) com respeito a $\mathbf{u}_1$. Essa deve ser uma otimização restrita para prevenir que $||\mathbf{u}_1|| \to \infty$. A restrição apropriada consiste na condição de normalização $\mathbf{u}_1^T\mathbf{u}_1 = 1$. Para reforçar essa restrição, pode-se introduzir um multiplicador de Lagrange (denotado por λ_1), e realizar a maximização irrestrita de:

$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u}_1). \tag{3.5}$$

Para encontrar a maximização da variância, pode-se igualar a derivada da função acima com respeito a $\mathbf{u}_1$ a zero. Com isso, pode-se observar que um ponto estacionário é obtido quando:

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1, \tag{3.6}$$

o que indica que $\mathbf{u}_1$ deve ser o autovetor de $\mathbf{S}$. Se multiplicarmos a equação acima por $\mathbf{u}_1^T$ e fizermos uso da restrição $\mathbf{u}_1^T\mathbf{u}_1 = 1$, pode-se observar que a variância é dada por:

$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \lambda_1. \tag{3.7}$$

Assim a variância será máxima quando $\mathbf{u}_1$ for igual ao autovetor que tiver o maior autovalor. Este será o primeiro componente principal.

Podem-se definir componentes principais adicionais de forma incremental escolhendo-se cada nova direção que maximize a variância dos dados nela projetados dentre todas as possíveis direções ortogonais àquelas já determinadas. Se for considerado o caso geral de projeção em um espaço d-dimensional, teremos d autovetores $\mathbf{u}$ da matriz de covariância $\mathbf{S}$ correspondentes aos D maiores autovalores $\lambda_1, \lambda_2, ..., \lambda_D$.

A título de exemplificação, considere o conjunto de dados $\mathbf{x}_n$ mostrado na Figura 3.8. Cada ponto representando os dados $\mathbf{x}_n$ com coordenadas x_1 e x_2 foi gerado da seguinte forma:

- Inicialmente, foram gerados dois vetores com 100 valores aleatórios normalmente distribuídos com médias 1 e 2 e desvios padrão 2 e 0.5, respectivamente;

- Esses vetores foram rotacionados -30 graus em relação ao eixo das abscissas ;

- Os valores das novas abscissas dos pontos foram atribuídos à variável x_1 e os valores das novas ordenadas dos pontos foram atribuídos à variável x_2.

A matriz de rotação aplicada aos dados é mostrada abaixo:

$$R = \begin{pmatrix} cos(-30°) & sen(-30°) \\ -sen(-30°) & cos(-30°) \end{pmatrix} \cong \begin{pmatrix} 0,866 & -0,5 \\ 0,5 & 0,866 \end{pmatrix} \tag{3.8}$$

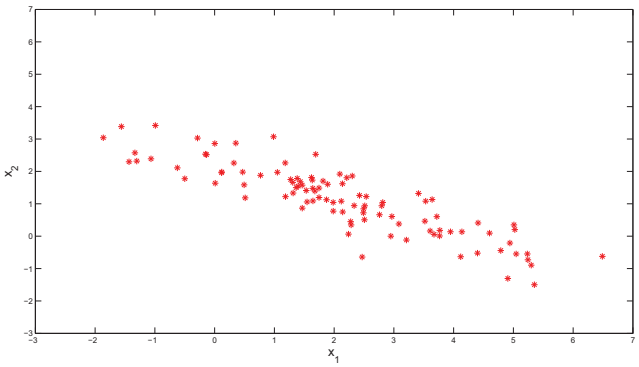


Figura 3.8: Dados aleatórios rotacionados em -30 graus.

Após resolver a Equação 3.6 para as matrizes de covariâncias de $\mathbf{x}_1 - \bar{x}_1$ e $\mathbf{x}_2 - \bar{x}_2$, obtém-se como autovetores os vetores $\mathbf{u}_1 = (0,863, -0.505)$ e $\mathbf{u}_2 = (0,505, 0,863)$, que correspondem aproximadamente aos valores das linhas das matrizes da Equação 3.8. A Figura 3.9 mostra a direção dos vetores $\mathbf{u}_1$ em magenta e $\mathbf{u}_2$ em azul. Isso significa que a PCA, nesse caso, pode ser interpretada como uma rotação do espaço, de forma que o primeiro componente principal represente o eixo de maior variância dos dados nele projetados e o segundo componente principal represente o eixo de menor variância de dados nele projetados. $\mathbf{u_1}$ é um vetor unitário com ângulo -30 graus no sistema de referências original e $\mathbf{u_2}$ é perpendicular a ele (possui ângulo 60 graus).

A próxima seção mostra como utilizar os dados de dimensionalidade reduzida por PCA para classificação de padrões.

3.3.2 Redes Neurais Artificiais

De acordo com Bishop [4], o termo rede neural artificial (RNA) tem suas origens na tentativa de encontrar uma representação matemática do processamento de informação em sistemas biológicos. Uma rede neural é um módulo matemático não linear e adaptativo que consiste de simples elementos chamados neurônios operando em paralelo e se comunicando através de conexões ponderadas. Numa analogia a sistemas biológicos, pode-se dizer que os neurônios da RNA, assim como os neurônios do sistema nervoso, possuem conexões de entrada (dendritos) e de saída (axônios).

Considere um vetor de variáveis de entrada $x_1, x_2, ..., x_D$. Cada neurônio corresponde matematicamente a combinação linear dessas variáveis, na forma:

$$\rho = \sum_{i=1}^{D} w_i x_i + b, \tag{3.9}$$

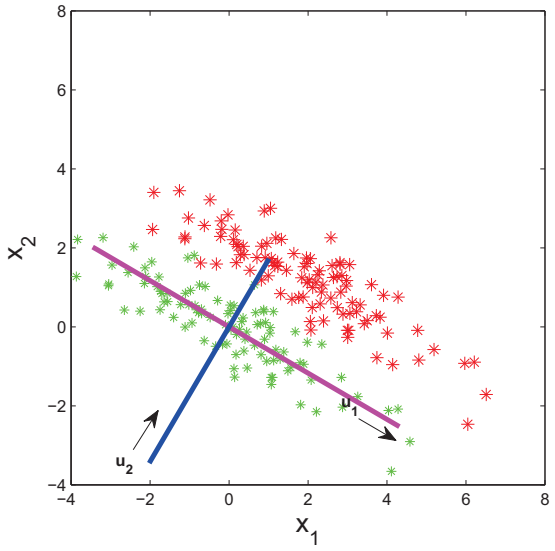


Figura 3.9: Em verde, os pontos da Figura 3.8 subtraídos das médias $\bar{x}_1$ e $\bar{x}_2$. $\mathbf{u_1}$ e $\mathbf{u_2}$ são os componentes principais dos dados, cujas direções são representadas pelas retas em magenta e azul, respectivamente.

onde os parâmetros w_i serão chamados pesos e o parâmetro b será chamado viés. Esses parâmetros devem ser determinados através de treinamento. Uma rede neural com um único neurônio e duas variáveis de entrada seria representada graficamente pelo esquema da Figura 3.10.

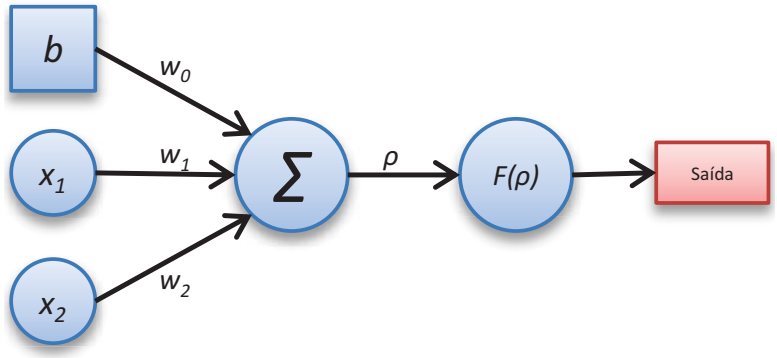


Figura 3.10: Típico modelo de um neurônio artificial.

Assim como os neurônios biológicos, os neurônios das RNAs possuem uma forma de processamento interno que gera um sinal de saída em função de um sinal de entrada. Para obtenção desse sinal de saída, o valor de ρ (chamado de "ativação") é transformado por uma função de ativação F, utilizada para restringir a sua amplitude. As funções de ativação definem a forma da saída dos neurônios. Várias são as funções de ativação que podem ser utilizadas para esse propósito (por exemplo, de Limiar, Linear por Partes, Sigmoide, Identidade), sendo que os intervalos de saída normalmente são definidos entre -1 e 1, 0 e 1 ou -0.5 e 0.5. Em problemas de classificação em múltiplas classes, cada valor de ativação pode ser transformado usando a função sigmoide tangente

hiperbólica:

$$F(\rho) = \frac{2}{2 + exp(-2\rho)} - 1 \tag{3.10}$$

cuja representação gráfica pode ser vista em 3.11.

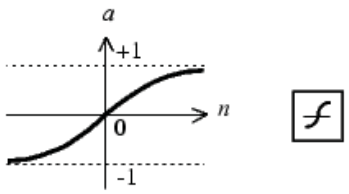


Figura 3.11: Função de ativação utilizada em redes neurais treinadas para reconhecimento de padrões.

O viés b da Figura 3.10 tem a função de diminuir ou aumentar o valor a ser processado pela função de ativação. Possui o efeito de transladar a função de ativação em relação à origem. Para o caso particular de duas variáveis de entrada, um neurônio sem o viés é como uma equação da reta sem o termo independente.

RNAs Multicamadas

As RNAs multicamadas são arquiteturas onde os neurônios são organizados em duas ou mais camadas de processamento [84]. As RNAs com apenas duas camadas são constituídas de uma camada de entrada que se conecta a uma camada de neurônios de saída. Os neurônios da camada de entrada são neurônios especiais, cujo papel é exclusivamente distribuir cada uma das entradas da rede (sem modificá-las) a todos os neurônios da camada seguinte. A forma mais simples deste tipo de rede consiste de um único neurônio na camada de saída, sendo conhecido como *perceptron*.

Conforme Haykin [25], o *perceptron* foi objeto de intensa pesquisa durante os anos 50 e 60, mas em 1969, M. Minsky e S. Papert provaram matematicamente que este tipo de estrutura de processamento apresenta limitações importantes e só pode ser aplicada com sucesso a uma classe muito restrita de problemas [48]. Mais especificamente foi provado que o perceptron é capaz de resolver apenas problemas linearmente separáveis.

No entanto, com a utilização de redes de múltiplas camadas, com pelo menos uma camada oculta (camada que não é nem entrada, nem saída), muitas das limitações apresentadas pelo *perceptron* deixam de existir, tornando as RNAs úteis em classes de problemas com soluções não lineares. Esta implementação, cuja arquitetura pode ser vista na Figura 3.12, recebeu o nome de *Multilayer Perceptron* (MLP). A Figura 3.12 mostra uma arquitetura típica de uma rede neural do tipo *feedforward*. Em redes desse tipo, as informações são propagadas através das entradas e saídas de cada neurônio, da camanda anterior para a camada posterior, sem retroalimentação de informação.

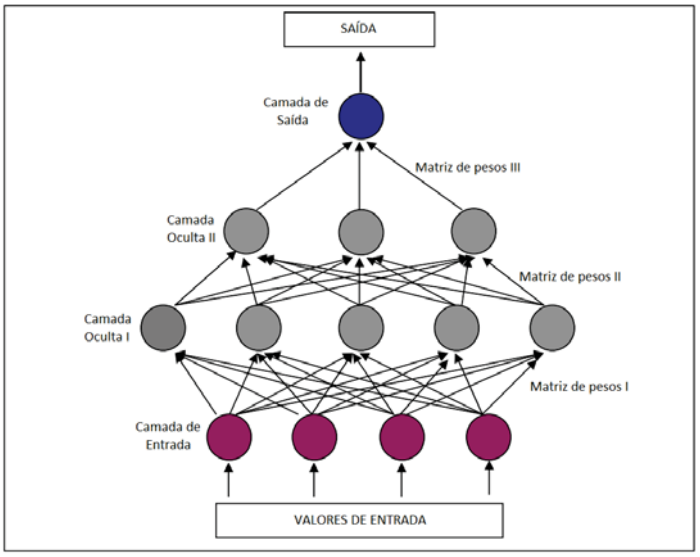


Figura 3.12: Arquitetura típica de uma rede neural *feedforward*.

Treinamento de RNAs

Conforme Wu & McLarty [84], a ideia fundamental por trás do treinamento de uma RNA é atribuir valores a um conjunto de pesos (inicializado normalmente de forma aleatória), aplicar os padrões à rede e verificar como ela responde a esses pesos. Se não responder de forma satisfatória, os pesos devem ser modificados por algum tipo de algoritmo (específico de cada arquitetura) e o processo deve ser repetido. Esta iteração deve continuar até que um critério de parada pré-definido seja atingido.

Cada passagem de treinamento sobre todos os padrões é chamada de época. Alterações nos pesos podem ser feitas a cada padrão processado ou após uma época inteira. Normalmente, os pesos são modificados após cada época.

O objetivo do treinamento é gerar uma rede com pesos que melhor atendam aos padrões recebidos de forma a generalizar a solução obtida para dados que venham a ser submetidos posteriormente. Para que isto aconteça, é necessário que seja evitada a situação de *overtraining* onde a rede "memoriza" os padrões recebidos e perde o seu poder de generalização. Para tanto, juntamente com a base de dados de treinamento, é processada uma base de validação que não altera os pesos da rede mas avalia o que foi aprendido até o momento. O ponto no qual a rede (com seus devidos pesos) deve ser armazenada é quando o erro da validação começa a subir.

O algoritmo de aprendizado é um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de RNAs. Estes algoritmos diferem entre si principalmente pelo modo como os pesos da rede são ajustados.

O algoritmo de aprendizado *backpropagation* é um método de aprendizado supervisionado para redes *feed-forward* de múltiplas camadas (MLP) proposto em 1986 por Rumelhart e colegas [61]. Nesse algoritmo, o procedimento de aprendizagem utiliza vetores de pesos que mapeiam um conjunto

de entradas para um conjunto de saídas. O aprendizado é realizado por ajuste iterativo dos vetores de pesos da rede para minimizar as diferenças entre a saída obtida e a saída desejada. Primeiramente, a rede é inicializada com pesos aleatórios. Durante o treinamento, um vetor de entrada é apresentado para a rede que determina os valores da saída. A seguir, o vetor de saída produzido pela rede é comparado com o vetor de saída esperado, associado ao vetor de entrada. A diferença entre os dois resulta num sinal de erro, que é retropropagado através da rede para o ajuste dos pesos. Este processo é repetido até que a rede responda, para cada vetor de entrada, com um vetor de saída com valores suficientemente próximos dos valores desejados.

Segundo Tissot e colegas [76], *Backpropagation* é um método baseado em descida de gradiente, o que significa que este algoritmo não garante que seja encontrado um mínimo global e pode estagnar em soluções de mínimos locais, onde ficaria preso indefinidamente. Contudo, é muito popular e amplamente utilizado no treinamento de RNAs.

Nesse trabalho, foi utilizado o *software* Matlab da *MathWorks Inc.*[4] para construção, treinamento e utilização das redes neurais *feed-forward* com algoritmo de treinamento *backpropagation* para reconhecimento de padrões. A função tangente hiperbólica da Equação 3.10 foi utilizada como função de ativação na camada oculta e uma função linear foi utilizada com ativação para a camada de saída. Foram treinadas RNAs para reconhecimento de unidades de ação ou emoções correspondentes aos dados rastreados pelo *Live Driver*, com redução de dimensionalidade feita por análise de componentes principais.

A próxima seção apresenta alguns dos parâmetros utilizados para guiar a animação de faces de avatares.

3.4 Parâmetros para Animação Facial

Sistemas de animação facial necessitam como dados de entrada uma série de parâmetros que indicam o grau de deformação de conjuntos de vértices que controlam regiões específicas da face. Atualmente, é muito utilizado o padrão MPEG-4 (*Moving Picture Experts Group*) para animação facial [53]. Esse padrão foi desenvolvido com base no Sistema de Codificação de Ações Faciais de Ekman et al. [18], exposto na Seção 3.1. O padrão MPEG-4 propriamente dito será mostrado na Seção 3.4.1, enquanto a Seção 3.4.2 mostra outra forma de parametrização para animações faciais.

3.4.1 O padrão MPEG-4 para Animação Facial

O padrão MPEG-4 de Animação Facial [53] especifica um conjunto de 84 pontos característicos (*Feature Points - FPs*) localizados na face. A Figura 3.13 mostra esses pontos característicos. Um subconjunto desses pontos atua como pontos de controle para os 68 parâmetros de animação (*Facial Animation Parameters* – FAPs), também definidos pelo padrão. Os dois primeiros FAPs descrevem ações em alto nível (6 expressões faciais e 14 visemas) e os restantes lidam com regiões específicas da face, descrevendo ações de mais baixo nível como "levantar o canto direito dos lábio" e "fechar a

[4] `http://www.mathworks.com/products/matlab/`

pálpebra superior esquerda". Os FAPs são codificados como valores numéricos, que são normalizados por um conjunto de unidades baseadas nas distâncias entre alguns pontos característicos principais da face, chamadas FAPU (*Facial Animation Parameter Units*). A Figura 3.14 apresenta essas distâncias e unidades. Com essa normalização, é possível animar faces com diferentes tamanhos, proporções e número de polígonos.

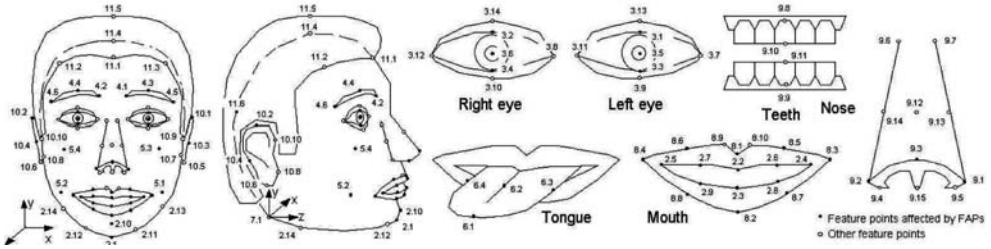


Figura 3.13: Pontos Característicos do padrão MPEG–4, retirada de [53].

Para a determinação da FAPU, a face deve estar em estado "neutro". Gerar uma animação baseada em FAPs consiste em prover, para cada frame de animação, a variação dos valores dos FAPs. Para cada frame, tem-se então uma sequência de valores dos FAPs, que é processada pela aplicação para gerar as animações na face e pode ser enviada pela rede ou salva em arquivos para posterior animação. É importante ressaltar que o padrão MPEG-4 apenas sugere os parâmetros envolvidos na animação de faces e não os métodos para deformá-las. Tendo um conjunto de valores de parâmetros, é necessário deformar os vértices da face para produzir a animação. Por exemplo, no padrão MPEG-4, cada FAP atua sobre um FP, que por sua vez influencia os vértices de sua vizinhança (que dependem da topologia da face e não são especificados pelo padrão), produzindo uma deformação na malha poligonal. Isso significa que cada valor de FAP é escalado pela sua FAPU para se obter o deslocamento do seu ponto de controle FP, e os vértices de sua zona de influência podem ser deformados pela aplicação de diferentes técnicas. A Figura 3.15 (à direita) mostra a atuação do FAP "Fechar Pálpebra Superior Esquerda", utilizando-se uma função gaussiana sobre o FP e sua vizinhança.

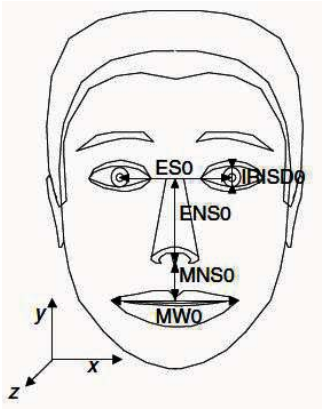


Figura 3.14: Medidas de distâncias que são utilizadas como unidades de animação no padrão MPEG-4.

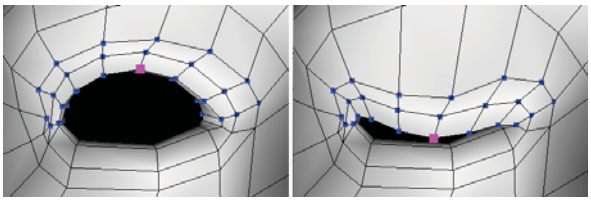


Figura 3.15: Exemplo de deformação na pálpebra da face 3D, mostrando o ponto de controle (FP, em rosa) e a sua zona de influência (pontos azuis), que sofre a deformação [59].

O sistema de animação do Laboratório de Humanos Virtuais da PUCRS utiliza-se de uma máscara de controle baseada no padrão MPEG4 para parametrização da animação facial de avatares.

3.4.2 *Blendshapes*

Blendshapes são matrizes de escalares que representam o grau de combinação entre duas formas do mesmo modelo 3D. Por exemplo, no caso específico de animação facial, considere uma face com a expressão neutra e outra face com um olho fechado, conforme mostrado na linha superior da Figura 3.16. Entre essas duas expressões extremas, existe uma série de passos de animação que são representados por um peso que indica quão próximo da face neutra está a face modelada (peso de *blendshape* com valor baixo) ou do sorriso aberto (peso de *blendshape* com valor alto). Em geral, artistas modelam um número n de expressões chave como "fechar olho direito", "fechar olho esquerdo", "abrir a boca", "franzir o nariz" e assim por diante, para um modelo com número de vértices V. Em um dado quadro, um vetor de n de pesos de *blendshapes* é transmitido ao sistema de animação, de forma a especificar o quão fechado está o olho (0 - olho na posição neutra, 1 - olho fechado), o quão aberta está a boca, o quão franzido está o nariz. Dessa forma, de acordo com [8], uma malha facial B com qualquer expressão pode ser representada pela combinação linear das malhas i geradas pelos artistas, a partir de uma máscara com a expressão neutra B_0, de acordo com a equação:

$$B = B_0 + \sum_{i=1}^{n} \alpha_i B_i, \tag{3.11}$$

onde $\mathbf{a} = \{\alpha_1, ..., \alpha_n\}$ é o vetor de coeficientes de *blendshapes*.

A Figura 3.16 mostra algumas sequências de imagens com um modelo 3D cuja face apresenta expressão neutra, pouco alterada e com a máxima deformação em uma dada coluna na matriz de *blendshapes*.

De acordo com [60], *Blendshapes* apresentam deformações suaves durante a interpolação de expressões alvo pré-existentes que podem ter sido esculpidas por artistas ou adquiridas por meio de dispositivos de captura de alta resolução como *scanners* 3D. Na Figura 3.16, as imagens da coluna da esquerda mostram a expressão neutra de um modelo de face tridimensional. As imagens da direita das duas primeiras linhas mostram expressões alvo definidas por artistas. As imagens centrais mostram combinações lineares das expressões neutras e das expressões definidas pelo artista. Na linha inferior, pode-se observar a combinação linear da expressão piscar da linha superior e abrir a

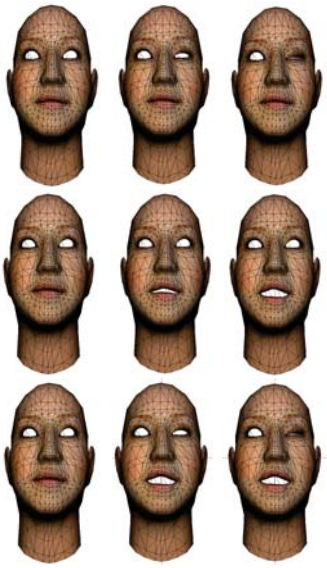


Figura 3.16: As imagens da esquerda mostram um modelo 3D com expressão neutra; as da esquerda mostram a deformação máxima do olho (acima) e da boca (no centro), conforme modelado por um artista; na coluna central estão a mistura das formas da coluna da direita com as da coluna da esquerda. Na sequência da linha inferior, mostra-se como se pode obter o *blendshapes* de três expressões: a face neutra, o olho piscando e a boca aberta.

boca da linha intermediária.

Uma abordagem muito utilizada para animação facial é definir as expressões alvo pelo Sistema de Codificação de Ações Faciais *Facial Action Code System - FACS* proposto por Paul Ekman et al. [19]. A Figura 3.17 mostra algumas dessas expressões alvo para o modelo com expressão neutra mostrada ao centro. Com combinações lineares dessas expressões por meio da Equação 3.11, aumenta-se indeterminadamente o número de expressões que o avatar pode assumir.

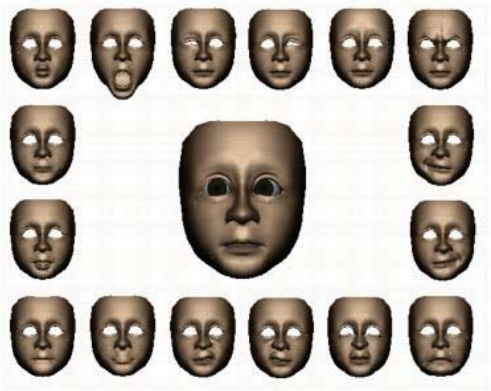


Figura 3.17: Exemplos de expressões faciais geradas por artistas para o mesmo modelo 3D. Um vetor de escalares pode representar o peso de cada expressão em determinado quadro da animação, gerando uma combinação de formas denominada *blendshapes* [45].

3.5 Banco de Dados Bosphorus

Os resultados de rastreamento dos modelos apresentados na Seção 3.2 são obviamente bidimensionais, dado que o processamento é feito em imagens de faces. Porém, os parâmetros de animação descritos na Seção 3.4 requerem informações em três dimensões. Além disso, a maioria desses métodos de rastreamento não reconhece as unidades de ação presentes em cada quadro rastreado (com exceção do *Live Driver*). Mesmo os parâmetros de animação informados pelo *Live Driver* devem ser pós-processados e reinterpretados pelos sistemas de animação, pois não são tão confiáveis quanto os pontos característicos rastreados em si.

A fim de se construir uma técnica que possa suprir essas duas carências - a falta de informações tridimensionais e rótulos das unidades de ação ou emoções - decidiu-se utilizar um banco de dados de faces tridimensionais digitalizadas em diferentes expressões e anotadas de acordo com o FACS. Assim sendo, realizou-se a busca por um banco de dados que satisfizesse os seguintes requisitos:

- fosse constituído de nuvens de pontos tridimensionais;
- apresentasse boa diversidade de indivíduos em diferentes expressões;
- fosse anotado de acordo com as unidades de ação descritas por Eckman (Seção 3.1), preferencialmente por pessoas certificadas.

Dados esses requisitos, o banco de dados escolhido foi o *Bosphorus Database* [64]. Esse banco de dados é composto por nuvens de pontos com informações de faces de 105 pessoas em até 35 expressões cada. Ao todo são 4666 nuvens de pontos, anotadas de acordo com o FACS. Um terço dessas pessoas são atores. Os mapas de profundidade são adquiridos usando um digitalizador 3D baseado em luz estruturada. Nesse digitalizador, a resolução nas dimensões horizontal, vertical e profundidade são respectivamente 0,3mm, 0,3mm e 0,4mm. Nesse tipo de sistema, os mapas de profundidade são adquiridos em uma única exposição, de forma que não há acúmulo de quadros como em outros sistemas. Isso se constitui numa vantagem, pois outros sistemas de aquisição (apesar de oferecerem maior resolução) requerem várias exposições ao longo de um intervalo de tempo relativamente longo, que dificulta a manutenção de uma expressão constante por parte do indivíduo rastreado. Simultaneamente à digitalização tridimensional, é adquirida uma fotografia em alta resolução (1600x1200 pixels) em formato *png*. Essas fotografias são disponibilizadas juntamente com as nuvens de pontos. Cada ponto do mapa de profundidades tem associado um ponto na respectiva imagem para que seja possível associar uma cor em RGB. Assim, é possível construir uma nuvem de pontos RGBD que pode ser armazenada em arquivos *ply* [5] por exemplo. Uma dessas nuvens de pontos pode ser visualizada na Figura 3.18. Para a aquisição desses dados, as pessoas são posicionadas a 1,5m aproximadamente do sensor em um quarto escuro, de forma a garantir a homogeneidade da iluminação.

[5] `http://www.mathworks.com/matlabcentral/fx_files/5459/1/content/ply.htm`, consultado em Julho de 2014.

A desvantagem desse sistema é a possibilidade de oclusão devido ao sombreamento de regiões da face dependendo do seu relevo, o que pode tornar a nuvem de pontos esparsa em algumas regiões, conforme pode ser visto na imagem central da Figura 3.18.

Além dessas informações, são disponibilizadas as coordenadas de 22 pontos de cada face tanto no mapa de profundidades como nas imagens, anotadas manualmente. Esses pontos podem ser vistos na imagem da direita da Figura 3.18.

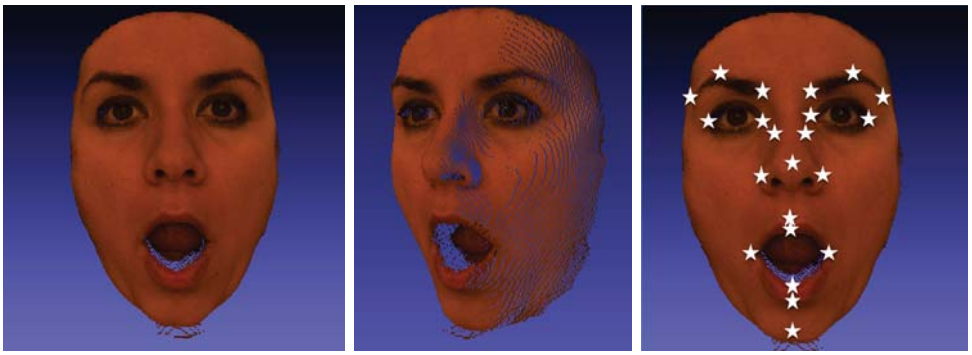


Figura 3.18: À esquerda, imagem da nuvem de pontos de uma pessoa executando a unidade de ação $LFAU_27$ da parte inferior da face. No centro, visão lateral da nuvem, mostrando regiões esparsas devido à oclusão. À direita, os pontos anotados manualmente, fornecidos com o banco de dados.

Conforme dito anteriormente, cada nuvem de pontos é anotada de acordo com o FACS (Seção 3.1). A Figura 3.19 ilustra como essa anotação é disponibilizada, juntamente com alguns exemplos de imagens com diferentes unidades de ação.

Figura 3.19: Anotação das expressões do banco de dados de acordo com o FACS. A codificação da anotação é apresentada à esquerda. As três imagens da direita mostram uma das pessoas cujas nuvens de ponto são fornecidas, realizando as unidades de ação "*Emotion - Happy*", "*Upper Facial Action Unit 1*" e "*Lower Facial Action Unit 27*".

Conforme a Figura 3.19, a anotação de uma nuvem de pontos contém as seguintes informações:

- identificador da pessoa cuja face foi digitalizada;
- categoria da expressão facial pode ser uma emoção (*Emotion*), uma unidade de ação da parte superior da face (UFAU - *Upper Face Action Unit*), uma unidade de ação da parte inferior da face (LFAU - *Lower Face Action Unit*) ou a indicação de que a expressão é neutra;
- identificação da expressão ou unidade de ação conforme o FACS;

- contador, indicando quantas réplicas da mesma unidade de ação ou emoção da mesma pessoa foram disponibilizadas.

Para a melhor compreensão desse trabalho faz-se necessário apresentar exemplos de cada unidade de ação ou emoção disponíveis no Bosphorus e utilizadas nesse trabalho. Para tanto, as Figuras 3.20 a 3.22 mostram um homem e uma mulher executando as mesmas unidades de ação, identificadas com a notação utilizada pelo Banco de dados Bosphorus.

As imagens da Figura 3.20 mostram as seis emoções cujas expressões faciais são universalmente reconhecíveis, segundo Ekman [18]. Essas expressões são combinações de unidades de ação que, em conjunto, podem ser reconhecidas como manifestação fisiológica de sentimentos.

Figura 3.20: Imagens de uma mulher e de um homem executando as expressões correspondentes aos sentimentos de raiva, asco, medo, alegria, tristeza e surpresa, com a notação usada no banco de dados Bosphorus correspondente.

As imagens da Figura 3.21, por sua vez, mostram as unidades de ação da parte superior da face (*Upper Face Action Units* - $UFAUs$). É possível perceber que as unidades de ação $UFAU\_1$ e $UFAU\_2$ referem-se a movimentos de sobrancelhas, enquanto as unidades de ação $UFAU\_43$ e $UFAU\_44$ são relacionadas ao movimento dos olhos. Já as imagens da Figura 3.22 mostram as unidades de ação relacionadas à parte inferior da face (*Lower Face Action Units* - $LFAUs$).

3.6 Análise de Procrustes

A análise de Procrustes determina a transformação linear - translação, reflexão, rotação e escala dos pontos da matriz Y que melhor a conforma aos pontos de uma matriz X [28]. A esse processo, dá-se o nome de registro dos pontos do conjunto Y com os pontos do conjunto X. De acordo com a mitologia Grega, Procrustes era o dono de uma estalagem que sujeitava seus hóspedes a

Figura 3.21: Imagens de uma mulher e de um homem executando as unidades de ação da parte superior da face (U_FAUs) com a notação usada no banco de dados Bosphorus correspondente.

medidas extremas para fazê-los caber em suas camas. Se os hóspedes fossem muito pequenos, ele os espichava; se eram muito altos, ele cortava suas pernas [28].

Suponha um conjunto de n pontos em um espaço Euclidiano q-dimensional, com coordenadas dadas pela matrix $n \times q$, X, que deve ser adaptada a outro conjunto em um espaço Euclidiano p-dimensional $(p \geq q)$ com coordenadas dadas pela matriz $n \times q$, Y. Assume-se que o k-ésimo ponto na matriz X tem uma correspondência com o k-ésimo ponto na matriz Y.

Primeiramente, $p - q$ colunas de zeros são adicionadas ao fim da matriz X a fim de que as matrizes fiquem com o mesmo número de dimensões.

Uma medida de discrepância entre as matrizes é então dada pela soma das diferenças quadradas, ϵ^2, entre os pontos correspondentes de X e Y, isto é:

$$\epsilon^2 = \sum_{k=1}^{n} (y_k - x_k)^T (y_k - x_k). \tag{3.12}$$

Os pontos da matriz X são escalados, transladados, rotacionados e refletidos para novas coordenadas $\mathbf{x}'$ dadas por:

$$\mathbf{x'_k} = s\mathbf{M}^T(x_k) + \mathbf{t}, \tag{3.13}$$

onde s é um fator de escala, $\mathbf{M}$ é a matriz de rotação e reflexão e $\mathbf{t}$ é o vetor de translação. Os valores ótimos desses elementos que minimizam o erro ϵ^2 são encontrados pelos procedimentos sumarizados a seguir:

- Deslocar os centroides dos dois conjuntos para a origem.
- Encontrar $\mathbf{M} = (\mathbf{x}^T\mathbf{y}\mathbf{y}^T\mathbf{x})$ e rotacionar X para $X\mathbf{M}$.
- Escalar a nova configuração de X multiplicando cada coordenada por $s = tr(\mathbf{x}^T\mathbf{y}\mathbf{y}^T\mathbf{x})/tr(\mathbf{x}^T\mathbf{x})$.

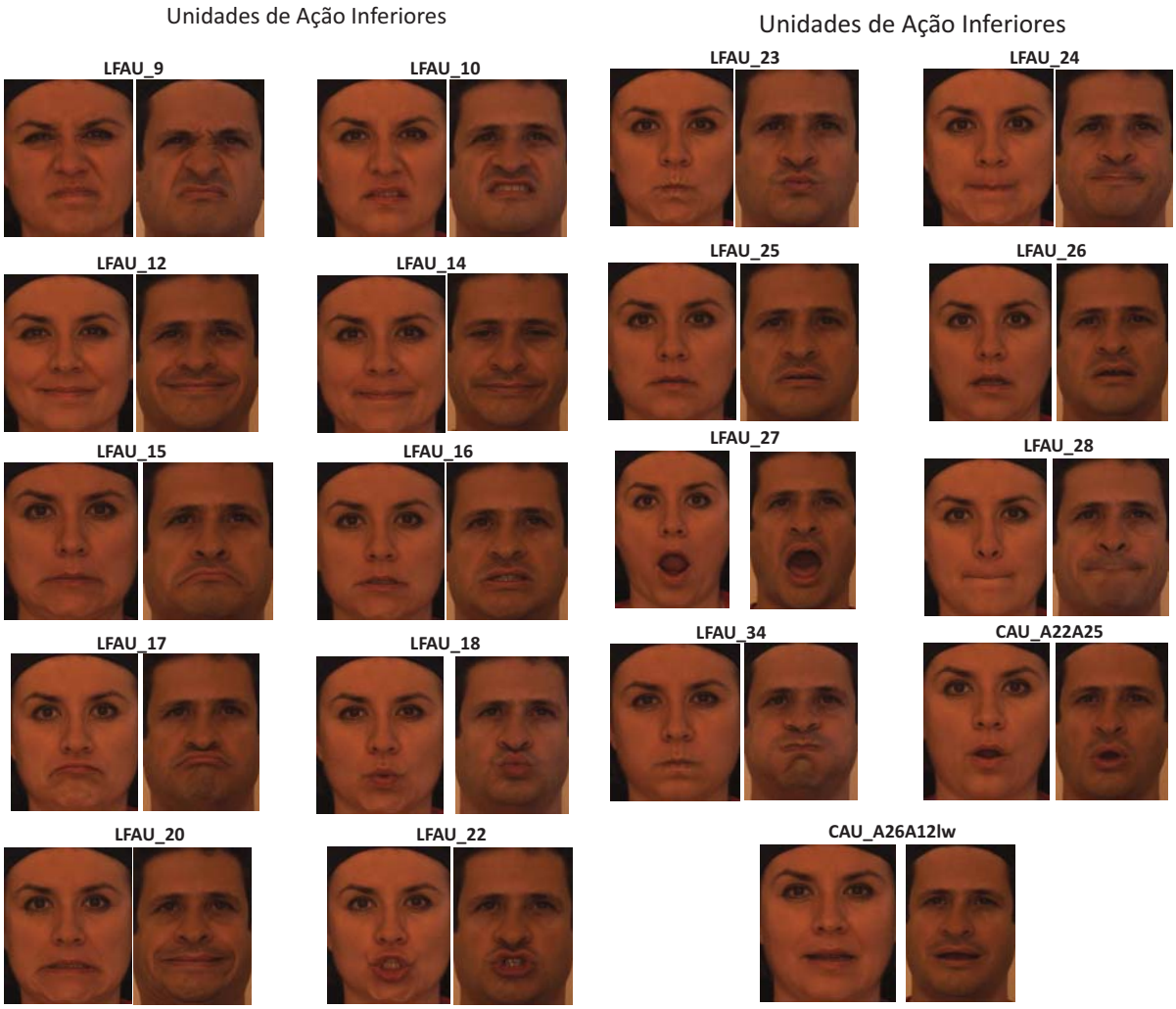


Figura 3.22: Imagens de uma mulher e de um homem executando as unidades de ação da parte inferior da face ($LFAUs$) com a notação usada no banco de dados Bosphorus correspondente.

- Calcular a estatística de Procrustes:

$$\epsilon^2 = 1 - \{tr(\mathbf{x}^T\mathbf{y}\mathbf{y}^T\mathbf{x})^{1/2}/tr(\mathbf{x}^T\mathbf{x})tr(\mathbf{y}^T\mathbf{y})\} \tag{3.14}$$

O valor de ϵ^2 pertence ao intervalo $[0, 1]$, com 0 significando um ajuste perfeito. A matriz M é obtida por decomposição em valores singulares de $\mathbf{x}^T\mathbf{y}$.

3.7 Distância de Hausdorff Modificada

Segundo [27], a distância de Hausdorff é a máxima distância de um conjunto ao ponto mais próximo de outro conjunto. Mais formalmente, a distância de Hausdorff, h, do conjunto A ao conjunto B é uma função de máximo/mínimo definida por:

$$h(A, B) = \max_{a \in A}\{\min_{b \in B}\{d(a, b)\}\}, \tag{3.15}$$

onde $d(a,b)$ é a distância Euclidiana entre os pontos $a \in A$ e $b \in B$. A Equação 3.15 encontra a mínima distância de cada ponto de A a todos os pontos de B e então seleciona o maior entre esses valores.

Dubuisson e Jain [17] propuseram a Distância de Hausdorff Modificada (DHM), que se mostrou mais eficiente para comparação de objetos em imagens do que a distância de Hausdorff original. A DHM H consiste na média entre os mínimos valores de distância dos pontos do conjunto A aos pontos do conjunto B e é dada pela equação:

$$H(A,B) = \frac{1}{N_A} \sum_{a \in A} (\min_{b \in B} \{d(a,b)\}), \tag{3.16}$$

onde N_A é o número de elementos de A. A DHM foi utilizada como métrica de semelhança entre conjuntos de pontos nesse trabalho.

Dados os conceitos fundamentais elencados nesse capítulo, apresenta-se agora o modelo proposto para aprendizado e utilização da Persona de um ator para animação de avatares.

4. MODELO PROPOSTO

Esse capítulo descreve o modelo proposto para solução do problema apresentado no capítulo de introdução: aprendizagem e utilização do estilo de movimento facial de um indivíduo ator A em um avatar dirigido pela performance de outro indivíduo usuário U, filmado por uma câmera monocular. A abordagem proposta baseia-se na construção de uma estrutura de dados representando o estilo de movimento do ator, chamada Persona. Em termos práticos, a Persona de um ator A é um conjunto de máscaras de controle de animação, que representam a forma como esse ator realiza cada unidade de ação ou emoção da face (exemplificadas nas Figuras 3.20 a 3.22). Tais unidades de ação ou emoção devem ser passíveis de serem reconhecidas automaticamente a partir de dados de rastreamento de componentes faciais, obtidos por *softwares* que operam em tempo real. No caso desse trabalho, a ferramenta escolhida para esse rastreamento foi o *Live Driver*, descrito na Seção 3.2.1.

Conforme mencionado, é necessário o reconhecimento automático das unidades de ação ou emoção realizadas pelo ator para associar uma classificação a cada máscara de controle incorporada à Persona de A. Essa classificação foi feita utilizando-se Análise de Componentes Principais e Redes Neurais (introduzidas na Seção 3.3.2), treinadas a partir de informações providas pelo banco de dados Bosphorus, descrito na Seção 3.5. Esse mesmo classificador é aplicado aos dados rastreados no rosto do usuário U. Assim, quando for reconhecida uma dada unidade de ação ou emoção no rosto de U (sorriso, por exemplo), o sistema deverá verificar na Persona do ator A qual a máscara de controle com a mesma classificação que melhor se adequa à expressão que o usuário U está realizando. Dessa forma, no caso do exemplo, quando o avatar for animado, ele irá sorrir da mesma forma que o ator A, ainda que seguindo o estímulo dado pelo usuário U.

O modelo desenvolvido para solução do problema de pesquisa desta tese divide-se em três etapas principais: **pré-processamento**, **construção da estrutura de dados Persona** e sua posterior **utilização**.

A etapa de *pré-processamento* está representada esquematicamente na Figura 4.1. Resumidamente, utilizam-se, nessa etapa, informações do Bosphorus e do rastreamento feito pelo *Live Driver* nas imagens desse banco de dados para que sejam obtidas máscaras 3D. Essas máscaras têm pontos de controle de animação adaptados a cada nuvem de pontos e a unidade de ação ou emoção correspondente é associada a cada uma delas. Os pontos de controle dessas máscaras são, então, submetidos à análise de componentes principais (PCA). O resultado da PCA é utilizado no treinamento supervisionado de redes neurais artificiais, que serão os classificadores automáticos de unidades de ação ou emoção de faces. Detalhes da etapa de pré-processamento, incluindo os processos 1 a 8 representados na Figura 4.1 são apresentados na Seção 4.1. Em suma, ao final dessa etapa, são obtidos os seguintes dados, destacados em vermelho na Figura 4.1:

- representações das expressões, via PCA;

- classificadores automáticos de unidades de ação ou emoção (RNAs);
- máscaras de controle de animação adaptadas às nuvens de pontos.

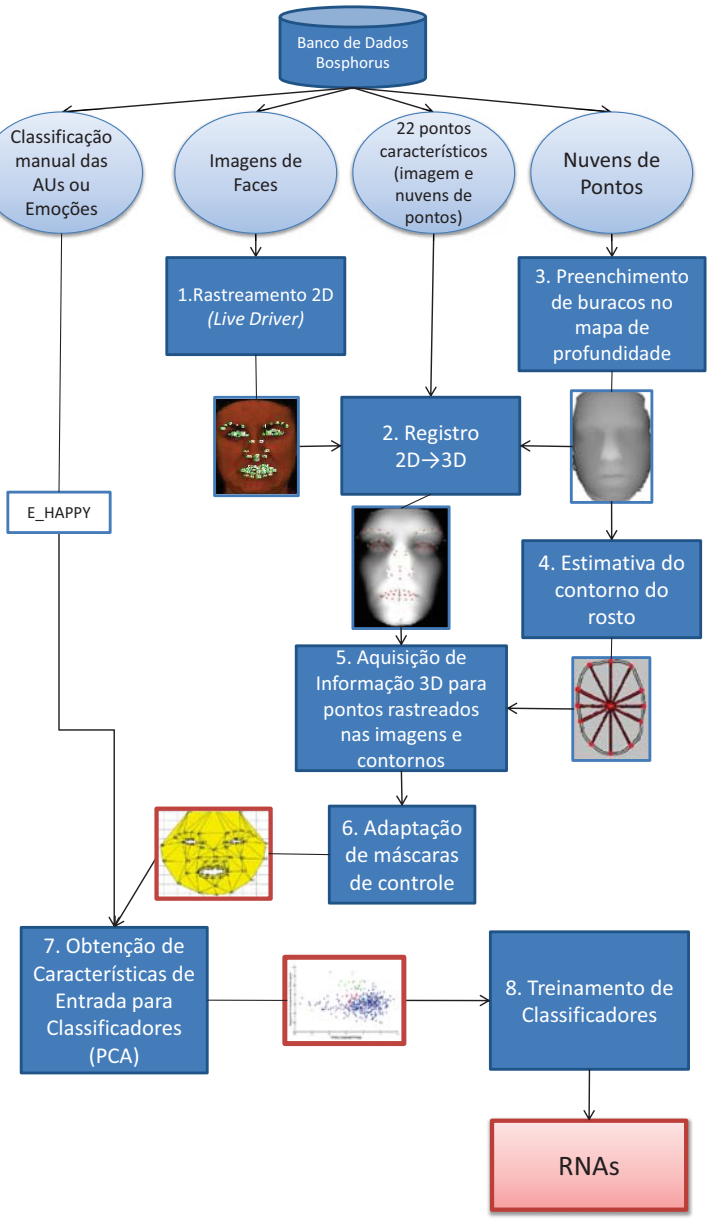


Figura 4.1: Esquema ilustrando a etapa de pré-processamento. Os dados que serão utilizados nas etapas de construção e utilização da persona estão destacados nos quadros em vermelho.

O processo de *Construção da Persona* de um ator A está esquematizado na Figura 4.2. Nesse esquema, observa-se que os dados rastreados no vídeo do ator são utilizados para obtenção da geometria 2D dos parâmetros de animação e como entrada para os classificadores. Tais classificadores foram treinados a partir de informações do banco de dados Bosphorus na etapa de pré-processamento. As máscaras com parâmetros de animação também geradas na estapa de pré-processamento são utilizadas para inferência de deslocamentos no plano transverso da face (inferência sobre a ter-

ceira coordenada dos parâmetros de controle). Mais detalhes sobre essa figura serão discutidos nas Seção 4.2.

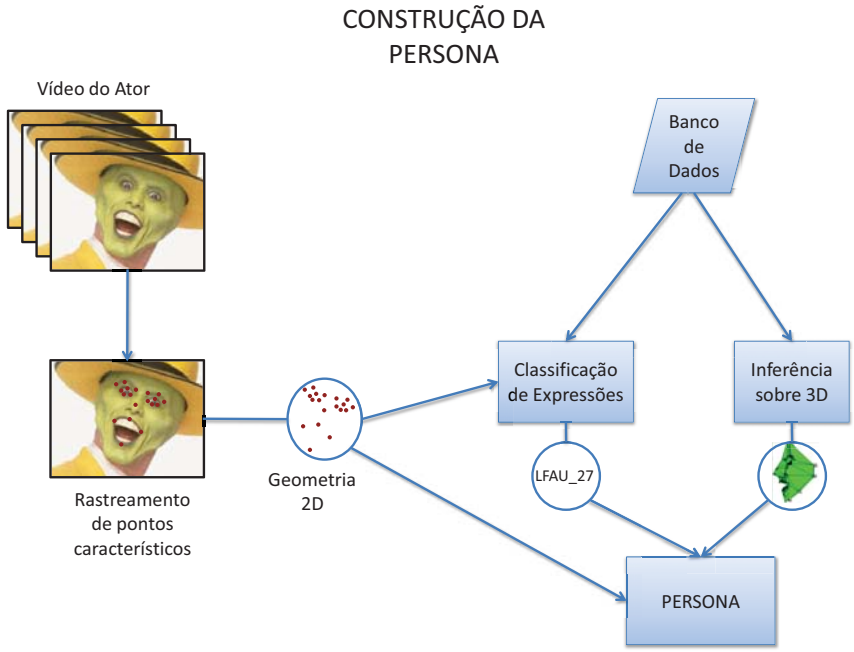


Figura 4.2: Esquema ilustrando a construção da Persona.

O processo de **utilização da Persona** tem como objetivo o provimento de parâmetros para animação do avatar de acordo com as unidades de ação ou expressões de um usuário qualquer U e com a Persona do ator A. Para isso, o movimento dos componentes faciais de U são rastreados pelo *Live Driver* e sua expressão ou unidades de ação são classificadas. Finalmente, deve-se escolher a máscara de controle da Persona de A adequada a essa classificação. A Figura 4.3 mostra um esquema ilustrando esse processo. Mais detalhes serão apresentados na Seção 4.3.

Decidiu-se que os requisitos do protótipo serão os seguintes:

- O usuário U deve ficar em frente à câmera com face paralela ao plano da imagem ou com pequeno ângulo de rotação - preferencialmente menor que 30 graus;
- Ao menos no primeiro segundo de captura (aproximadamente 30 quadros), o usuário U deve permanecer com uma expressão neutra;
- Serão utilizados, como dados de entrada, os pontos característicos fornecidos pela ferramenta de desenvolvimento de *software Live Driver* descrita na Seção 3.2.1;
- O banco de dados Bosphorus descrito na Seção 3.5 será utilizado para dois fins, conforme esquema das Figuras 4.2 e 4.3:
  - no treinamento supervisionado do classificador das expressões;
  - na inferência de informações sobre o deslocamento dos componentes faciais no plano transverso à face (vide Figura 4.2).

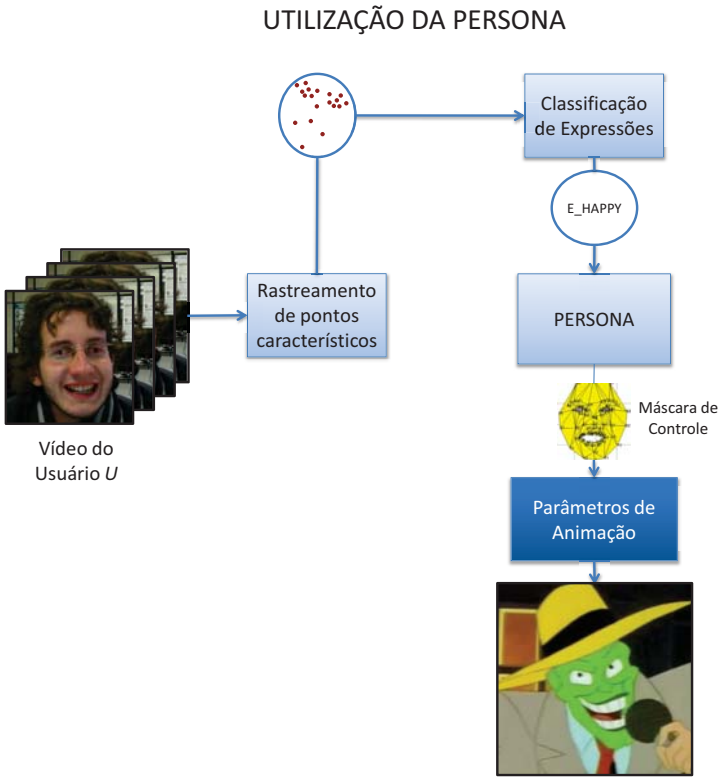


Figura 4.3: Esquema ilustrando a utilização da Persona.

Nesse trabalho foram utilizados os sistemas de coordenadas mostrados na Figura 4.4. As imagens superiores da Figura 4.4, com visualização dos planos coronal, tranverso e sagital da face, mostram o sistema de referências adotado para nuvens de pontos, máscaras de controle e modelos de face 3D. A imagem inferior mostra o sistema de referências utilizado para as coordenadas de imagem, com origem no canto superior esquerdo.

O restante deste capítulo está organizado de acordo com as etapas do método. A Seção 4.1 descreve em detalhe as etapas de pré-processamento. A Seção 4.2 explica o processo de criação da Persona. Já a Seção 4.3 apresenta uma metodologia para utilização da Persona criada.

4.1 Pré-Processamento

Conforme mencionado na Seção 3.5, o banco de dados Bosphorus provê, para cada nuvem de pontos tridimensional, uma imagem em formato *png*[1] correspondente e a classificação da unidade de ação ou emoção executada no momento da aquisição, de acordo com o FACS (Seção 3.1). O banco de dados provê, ainda, as seguintes informações:

- 3 coordenadas Euclidianas (x, y, z) de cada ponto digitalizado. Nesse sistema de coordenadas, o eixo x corresponde à coordenada horizontal, o eixo y à coordenada vertical e z à profundidade. Existem em torno de 35000 pontos em cada nuvem;

[1] *Portable Network Graphics*

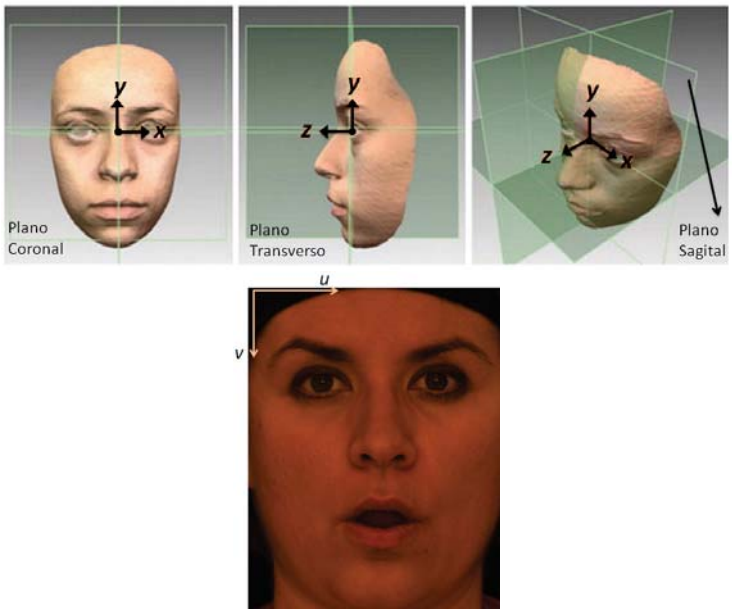


Figura 4.4: Sistemas de referência utilizados nesse trabalho. Acima, sistema de referência tridimensional utilizado em nuvens de pontos, máscaras de controle e modelos geométricos 3D, com visualização dos planos coronal (à esquerda), sagital (no centro) e transverso à face (à direita) . A imagem abaixo, mostra o sistema de referências para coordenadas de imagem.

- 2 coordenadas de imagem (u, v) indicando a posição na imagem *png* correspondente a cada ponto da nuvem;
- o menor valor de profundidade z_{min}, que corresponde ao plano de fundo da nuvem de pontos;
- número de linhas e colunas (n_l, n_c) da nuvem de pontos;
- coordenadas euclidianas tridimensionais de 22 pontos característicos da face anotados manualmente nas nuvens de pontos capturadas, aos quais denominou-se conjunto $\mathcal{B} = \{(x_{\mathcal{B}i}, y_{\mathcal{B}i}, z_{\mathcal{B}i})\}$ com $i = 1, 2, ..., 22$;
- coordenadas de imagem de 22 pontos característicos da face anotados manualmente nas imagens, ao qual denominou-se conjunto $\beta_I = \{(u_{\beta_i}, v_{\beta_i})\}$ e que podem ser vistos na imagem superior da Figura 4.5. Os índices dos pontos desse conjunto correspondem aos índices dos pontos do conjunto $\mathcal{B}$. Por exemplo, a ponta do nariz corresponde ao ponto $i = 57$ tanto no conjunto $\mathcal{B}$ como no conjunto β_I.

As próximas seções explicam como essas informações são utilizadas para adaptação de máscaras de controle de animação às nuvens de pontos.

4.1.1 Mapeamento dos Pontos Rastreados pelo *Live Driver* para a Nuvem de Pontos do Bosphorus

O primeiro passo na etapa de pré-processamento (conforme pode ser visto na Figura 4.1), é a obtenção dos pontos característicos rastreados pelo *Live Driver* nas imagens fornecidas no banco

de dados Bosphorus. Antes de tudo, entretanto, foram excluídas do treinamento nuvens de pontos com oclusões e rotações da face em relação ao plano da imagem maiores que 30 graus a fim de que se obtenha um conjunto de informações confiável e livre de artefatos para o treinamento.

Após, aplicou-se a ferramenta *Live Driver* às imagens fornecidas com o banco de dados Bosphorus. Como essa ferramenta foi desenvolvida para rastrear pontos característicos da face em sequências de vídeo, verificou-se que o resultado obtido não é muito acurado quando as imagens mudam bruscamente quadro a quadro. Provavelmente, isso decorre do uso de informação temporal da sequência de imagens que constitui o vídeo nos algoritmos de rastreamento. Essa suposição tem base no fato de que, quando ocorre mudança brusca de cena em uma sequência de vídeo, o rastreamento é prejudicado e somente passa a ocorrer com sucesso decorridos alguns novos quadros. Para que esse problema seja contornado, replicou-se 20 vezes cada imagem e gravaram-se somente as informações de rastreamento correspondente à vigésima imagem. Assim, para cada imagem, obteve-se um conjunto de 64 pontos Λ_I com coordenadas (u_Λ, v_Λ) detectados pelo *Live Driver*, conforme as imagens central e inferior da Figura 4.5. O índice I indica que esses pontos estão em coordenadas de imagem.

Em seguida, realizou-se o registro de cada conjunto β_I de pontos anotados nas imagens em relação às duas primeiras coordenadas $(x_{\mathcal{B}i}, y_{\mathcal{B}i})$ dos conjuntos $\mathcal{B}$ anotados nas nuvens de pontos[2]. Essa etapa de registro está representado na Figura 4.1 pelo processo número 2 (Registro 2D→3D). O objetivo desse procedimento de registro é encontrar a transformação rígida que, se aplicada aos conjuntos β_I, melhor os conforma às duas primeiras coordenadas dos pontos dos conjuntos $\mathcal{B}$. Posteriormente, pode-se estender essa transformação rígida a quaisquer pontos das coordenadas de imagem em pontos correspondentes da nuvem de pontos. Para isso, realizou-se a análise de Procrustes [30] descrita na Seção 3.6 a fim de que se obtenha uma matriz de rotação $\mathbf{R}$, um vetor de translação $\mathbf{t}$ e um escalar s correspondente a um fator de escala. Com esses dados, pode-se calcular as coordenadas bidimensionais na nuvem de pontos de cada conjunto Λ_I de 64 pontos rastreados pelo *Live Driver*. Cada ponto desse novo conjunto $\Lambda = \{(x_{\Lambda_k}, y_{\Lambda_k})\}$, com $k = 1, 2, ..., 64$, será obtido pela equação:

$$(x_{\Lambda_k}, y_{\Lambda_k}) = s.(u_{\Lambda_k}, v_{\Lambda_k}).\mathbf{R} + \mathbf{t}. \tag{4.1}$$

Após esse procedimento, é necessário atribuir valores à terceira coordenada de cada ponto do conjunto Λ de acordo com a nuvem de pontos correspondente. A próxima seção explica como se dá essa atribuição.

4.1.2 Atribuição de valores de profundidade aos pontos rastreados

Para facilitar a atribuição de valores de profundidade aos pontos rastreados, as informações de cada nuvem de pontos são representadas por meio de um mapa de profundidades que pode ser visualizado à esquerda da Figura 4.6. Para obtenção desse mapa, as coordenadas $(x,\ y)$ de cada ponto da nuvem passam a corresponder à coluna e à linha do mapa, respectivamente. A coordenada

[2] As informações para registro entre os conjuntos $\mathcal{B}$ e β_I não são fornecidas pelo Bosphorus.

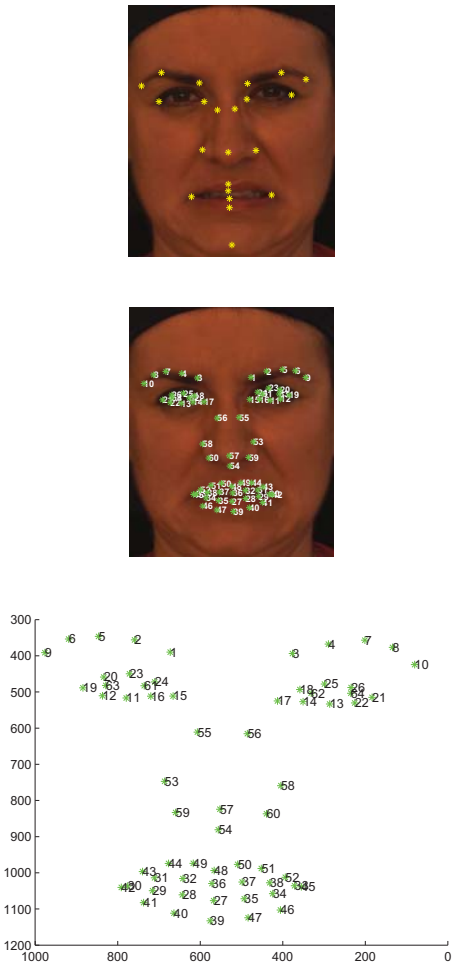


Figura 4.5: Acima, pontos manualmente anotados ($\beta_I = \{(u_{\beta_i}, v_{\beta_i})\}$), disponíveis no Bosphorus. No meio e abaixo, conjunto pontos Λ_I resultante do rastreamento da ferramenta *Live Driver* aplicada às imagens do Bosphorus.

z foi mapeada para um tom de cinza, para fins de visualização. Na imagem central da Figura 4.6, podem ser vistas coordenadas $x_{\mathcal{B}}$ e $y_{\mathcal{B}}$ do conjunto de pontos $\mathcal{B}$ em verde e os pontos do conjunto $\Lambda = \{(x_{\Lambda_k}, y_{\Lambda_k}\})$ após aplicada a transformação da Equação 4.1, em vermelho. A imagem da direita da Figura 4.6 corresponde ao mapa de profundidade da esquerda, adicionando-se a informação de cor. Esse mapa é conhecido como RGBD.

O valor de profundidade z_{Λ_k} deve ser obtido pela verificação do valor de profundidade do mapa que corresponde a cada coordenada bidimensional (x_k, y_k) dos pontos do conjunto Λ. Porém, esses pontos podem ficar sobre buracos do mapa de profundidades. Antes de que se atribua o valor da coordenada z_{Λ_k} , então, alguns reparos na nuvem de pontos devem ser realizados. Esses reparos correspondem ao preenchimento de buracos do mapa de profundidades e são discutidos na próxima seção.

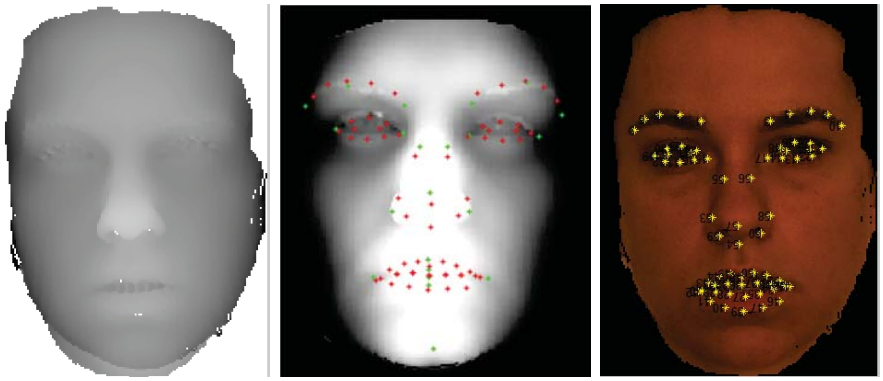


Figura 4.6: À esquerda, mapa de profundidades de uma nuvem de pontos. No centro podem ser vistos os pontos manualmente anotados em verde e os pontos do conjunto Λ em vermelho. À direita, o mapa RGBD com os pontos do conjunto Λ em amarelo.

Preenchimento de Buracos das nuvens de Pontos do Bosphorus

Como pode ser observado na imagem da esquerda da Figura 4.6 existem alguns "buracos" em regiões da face digitalizada dos indivíduos. Isso ocorre devido à oclusão, sombreamento, falhas de digitalização ou deficiências devidas à resolução em locais onde o relevo da face tem alto gradiente. Tais falhas foram salientadas em amarelo na nuvem de pontos da Figura 4.7.

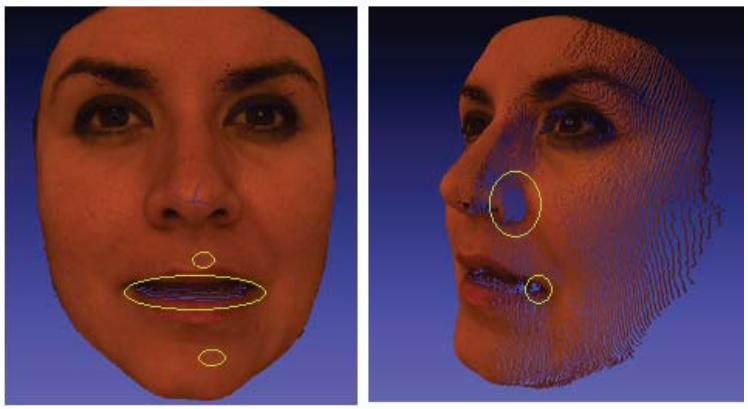


Figura 4.7: Buracos em nuvem de pontos do Bosphorus. A imagem da direita corresponde à rotação da mesma nuvem de pontos visualizada à esquerda.

Como pontos do conjunto Λ podem ficar sobre regiões com buracos, são necessários procedimentos para inferência da informação de profundidade nesses locais. Esse procedimento de "fechamento de buracos" (processo 3 na Figura 4.1) é feito da seguinte forma: em primeiro lugar, determinam-se os pontos correspondentes a buracos na nuvem e que estão dentro do poliedro imaginário cujos vértices da base são formados pelos pontos do conjunto Λ. Esses pontos possuem profundidade igual a z_{min}, ou seja, têm a mesma profundidade do plano de fundo das nuvens de pontos. Para cada um desses pontos, com coordenadas (r_j, s_j) no mapa de profundidades, determina-se as coordenadas de pontos da vizinhança: $\{(r_j+1, s_j),(r_j+1, s_j+1),(r_j, s_j+1),(r_j-1, s_j+1),(r_j-1, s_j),(r_j, s_j-1),(r_j, s_j-1),(r_j-1,s_j-1)\}$. Verifica-se então quais deles têm valor de profundidade maior que z_{min}. O valor da profundidade de cada ponto (r_j, s_j) corresponderá a média aritmética dos valores

z dessa vizinhança que forem maiores que z_{min}. Esse procedimento é realizado recursivamente, até que todos os buracos sejam preenchidos.

A Figura 4.8 ilustra esse processo. Considere que a grade da esquerda seja uma região de um hipotético mapa de profundidades. O valor numérico indica o valor da coordenada z_j em cada célula de coordenada (x_j, y_j). As células em cinza correspondem a buracos na nuvem de pontos com profundidade $z_j = z_{min}$. O ponto com coordenadas $x = 2$ e $y = 3$ terá atribuída a profundidade $z = 11$, que é a média aritmética dos valores da vizinhança que não correspondem a buracos, ilustrados em vermelho na imagem central dessa figura. Da mesma forma, a imagem da direita mostra a atribuição da profundidade do ponto de coordenadas $x = 2$ e $y = 2$.

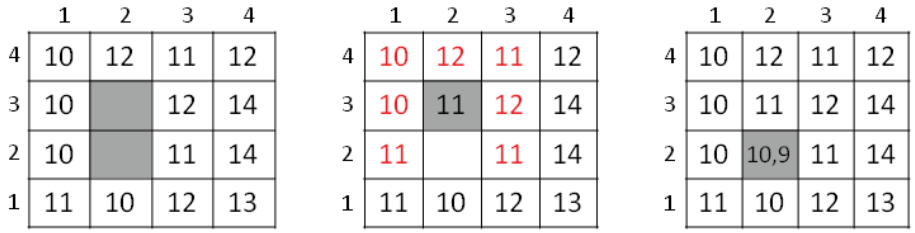


Figura 4.8: Ilustração de três passos sequenciais do procedimento para preenchimento de buracos nas nuvens de pontos

Após esse procedimento, o valor da profundidade z do mapa de profundidades é atribuído a cada um dos 64 pontos do conjunto Λ, que passam agora a ser tridimensionais. Assim são obtidos n conjuntos Λ, um para cada nuvem de pontos não eliminada pelos critérios mencionados no início da Seção 4.1.1.

4.1.3 Estimativa do Contorno do Rosto

Em um processo de animação facial, é essencial a informação do movimento de pontos relacionados ao queixo e à mandíbula. Entretanto, tais pontos não são providos pelo *Live Driver*, conforme ilustra a imagem central da Figura 4.5. Para a estimativa do contorno das faces do banco de dados Bosphorus (processo número 4 da Figura 4.1), assume-se que pertencem à face todos os pontos cuja coordenada z é maior que a profundidade $(z_{\Lambda_{57}} - \delta_1)$ (a ponta do nariz possui índice $k = 57$) e menor que $(z_{\Lambda_9} + \delta_2)$ (a extremidade externa da sobrancelha direita possui índice $k = 59$). Os valores δ_1 e δ_2 são distâncias empiricamente fixadas em $\delta_1 = 0,5cm$ e $\delta_2 = 1cm$. Com isso, excluem-se da nuvem de pontos dados pertencentes ao pescoço, parte superior da cabeça e orelhas. Assim, obtém-se uma imagem como a da esquerda da Figura 4.9. Após, procede-se uma binarização dessa imagem e determina-se o contorno usando detecção de bordas de Sobel [68]. Em seguida, esse contorno é dilatado utilizando-se como elemento estruturante uma matriz 3×3 de valores unitários, obtendo-se uma imagem conforme a imagem central da Figura 4.9. Determina-se então o centroide do contorno assim obtido, de coordenadas (x_c, y_c). A partir do centroide, a cada 22,5 graus é traçada uma reta e determina-se sua interceptação com o contorno, de forma a se obterem 16 pontos.

Esse ângulo foi fixado nesse valor pois boa parte dos 16 pontos encontrados coincidem com os 14 pontos no contorno na máscara de controle de animação de avatares utilizada no Laboratório de Simulação de Humanos Virtuais da PUCRS, mostrada na Seção 4.1.4. Os pontos da máscara de controle que não coincidem com os 16 pontos do contorno da nuvem de pontos determinados são calculados conforme explicado na Tabela 4.1 da Seção 4.1.4.

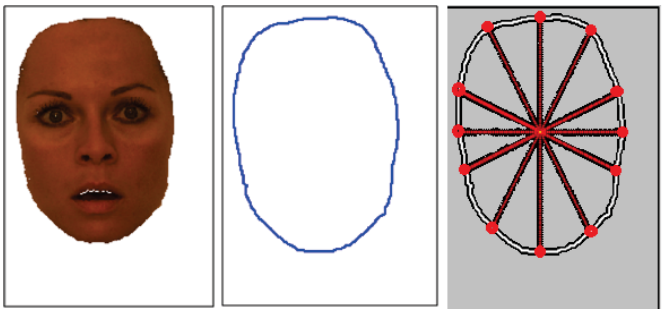


Figura 4.9: Processo para encontrar pontos do contorno da máscara. À esquerda, pontos considerados como pertencentes à face; no centro, contorno da face da imagem à esquerda; à direita, 16 pontos do contorno estimados.

Tendo sido obtidas as coordenadas bidimensionais dos 16 pontos pertencentes ao contorno da face (mostrados na imagem da direita da Figura 4.9), retorna-se ao mapa de profundidades para que seja estimada a terceira coordenada de cada um desses pontos. Assim, cada conjunto de pontos Λ é estendido para 80 pontos tridimensionais (64 rastreados pelo *Live Driver* e 16 do contorno da face), que constituem-se na saída do processo 5 da Figura 4.1.

Mesmo com todos os procedimentos acima listados, podem ainda haver erros na determinação de conjuntos Λ adequados, seja por falhas na detecção dos pontos característicos pelo $LiveDriver$, seja por lacunas grandes demais na nuvem de pontos, ou ainda por deficiências no processo de aquisição das nuvens de pontos do Bosporus, visto que a cor pode estar deslocada em relação à profundidade de algumas regiões. Tais problemas ocorrem devido ao alto gradiente de profundidade existente especialmente quando a boca não está fechada, o que ocasiona regiões esparsas na nuvem de pontos, conforme pode ser visualizado na Figura 4.7. Por isso, cada nuvem foi inspecionada individualmente. Com o objetivo de realizarem-se ajustes manuais, se necessário, foi desenvolvida uma ferramenta para ajuste manual de pontos. Assim, após todos esses procedimentos, obtém-se n conjuntos Λ de 80 pontos tridimensionais confiáveis. Após as inspeções manuais, foram obtidos $n = 930$ conjuntos Λ para o banco de dados Bosphorus.

4.1.4 Adaptação das Máscaras de controle 3D às Nuvens de Pontos

Conforme explicado na Seção 3.4, é possível animar um modelo geométrico de face por meio da informação do deslocamento de pontos de controle em relação à face neutra. Na equipe de animação de faces de avatares do Laboratório de Humanos Virtuais da PUCRS, foi definido que os pontos de controle para animação facial serão baseados no conjunto de FPs do padrão MPEG4, conforme Seção 3.4.1, os quais incluem os pontos rastreados pelo *Live Driver*. Esses pontos foram

triangulados de forma a criar máscaras de controle. Os triângulos são utilizados no processo de animação para a definição de associações dos vértices do modelo 3D da face do avatar aos pontos de controle. Por meio dessas associações, a máscara com 86 vértices e 130 polígonos controla o movimento de avatares com um número de polígonos que pode ser muito maior. Resultados desse processo de animação podem ser vistos no Capítulo 5. Imagens de uma máscara de controle padrão podem ser observadas na Figura 4.10.

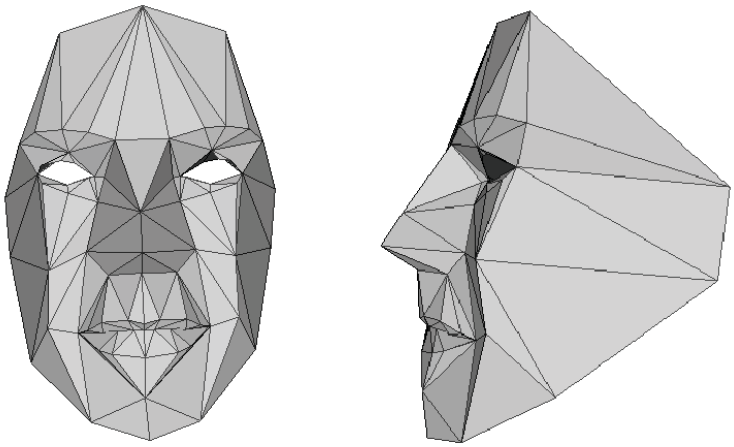


Figura 4.10: Duas vistas da máscara de controle padrão.

Conforme explicado nas seções anteriores, cada conjunto de dados Λ tem 80 pontos tridimensionais, correspondentes aos pontos rastreados pelo *Live Driver* com coordenada z obtidas nas nuvens de pontos do Bosphorus. Para que essas informações sejam úteis no restante desse trabalho, é necessário transformar esses pontos em máscaras de controle de animação.

A maior parte dos 80 pontos do conjunto Λ tem correspondência direta com os pontos da máscara de controle. As exceções a esse fato são os quatro pontos pertencentes às íris e alguns pontos do contorno e do interior da face. Os quatro pontos do conjunto Λ referentes às posições das íris serão desconsiderados nesse trabalho. Faltam, então, estimar as coordenadas de 12 pontos da máscara de controle no interior da face e 14 pontos do contorno não rastreados pelo *Live Driver* (adaptados dos 16 pontos encontrados automaticamente no contorno). Esses 26 pontos são mostrados na Figura 4.11.

Conforme se pode observar na imagem superior da Figura 4.11, esses pontos do interior da face correspondem semanticamente a locais que, em geral, têm pouca variação de textura. Isso torna difícil estimar sua localização por algoritmos de visão computacional. Por isso, a estimativa de sua localização nas nuvens de pontos do Bosphorus é feita com base em geometria, a partir dos pontos rastreados pelo *Live Driver*. Já os pontos do contorno da máscara de controle são determinados a partir das 16 coordenadas dos contornos das nuvens de pontos determinadas conforme a Seção 4.1.3. A Tabela 4.1 mostra como as coordenadas dos 26 vértices salientados na Figura 4.11 são calculadas. Nessa tabela, as letras da coluna da esquerda representam os pontos marcados em vermelho na imagem superior da Figura 4.11. Já os índices k das coordenadas x e y das colunas central e direita

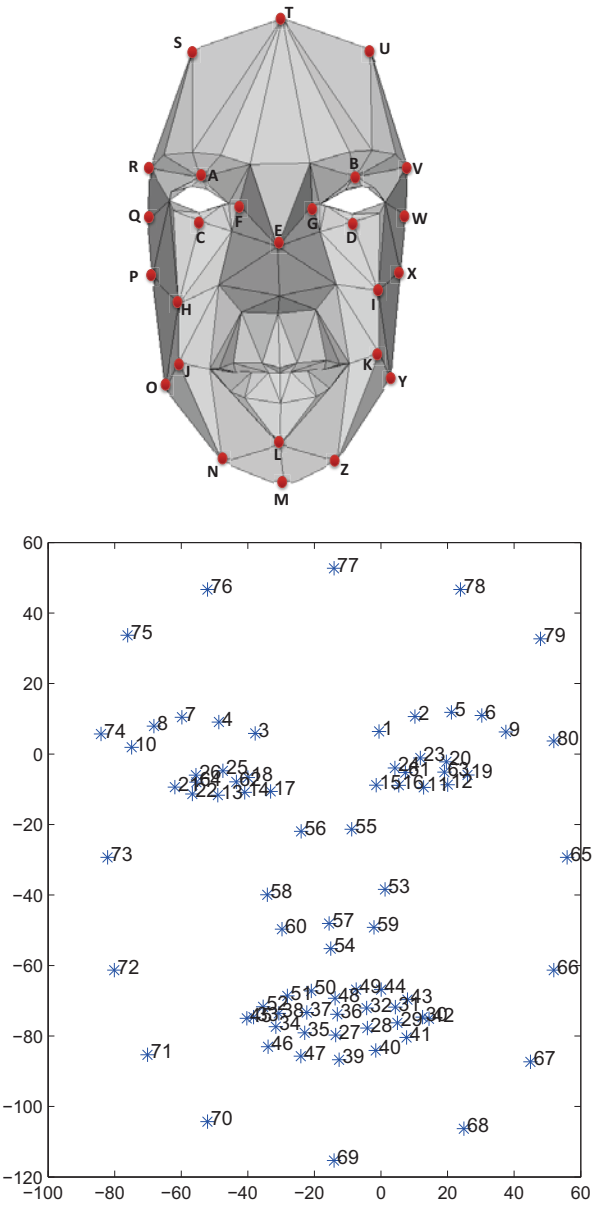


Figura 4.11: Acima, vértices da máscara de controle não rastreados pelo *Live Driver* marcados em vermelho. Abaixo, os pontos de um conjunto Λ com sua respectiva numeração.

referem-se aos índices dos pontos do conjunto Λ mostrados na imagem inferior da Figura 4.11. $x_{\mathcal{K}'}$ e $y_{\mathcal{K}'}$ serão, então, as coordenadas bidimensionais dos vértices da máscara de controle em vermelho. Note que não é utilizada a informação de profundidade das nuvens de pontos, pois esse mesmo conjunto de equações deve ser utilizado posteriormente para a geração das máscaras de controle a partir dos pontos rastreados pelo *Live Driver* em vídeos de atores ou usuários. Nesses casos, apenas informações bidimensionais estão disponíveis.

Dadas as coordenadas $(x_{\mathcal{K}'}, y_{\mathcal{K}'})$ determinadas pelas equações mostradas na Tabela 4.1, retorna-se aos mapas de profundidade para determinar as coordenadas $z_{\mathcal{K}'}$ correspondentes. Dessa forma, são obtidas n máscaras de controle formadas pela triangulação dos conjuntos de 86 vértices, as quais denominaremos $\mathcal{K}'$. Para relembrar, dos 64 pontos do rastreados pelo *Live Driver*, 4 são eliminados

Tabela 4.1: Atribuição de valores às coordenadas de vértices da máscara de controle não rastreados pelo *Live Driver*.

Ponto da Máscara $\mathcal{K}'$ (Figura 4.11)	$x_{\mathcal{K}'}$	$y_{\mathcal{K}'}$
A	$\frac{x_{64}+x_{62}}{2}$	$y_{62}+\frac{y_7-y_{62}}{2}$
B	$\frac{x_{61}+x_{63}}{2}$	$y_{61}+\frac{y_5-y_{61}}{2}$
C	x_{13}	$y_{13}-(y_{62}-y_{13})$
D	x_{11}	$y_{11}-(y_{61}-y_{11})$
E	$\frac{x_{56}+x_{55}}{2}$	$\frac{y_{56}+y_{55}}{2}$
F	x_{56}	y_{17}
G	x_{55}	y_{15}
H	x_{22}	y_{57}
I	x_{12}	y_{57}
J	x_{22}	y_{45}
K	x_{12}	y_{42}
L	x_{39}	$y_{39}-(y_{54}-y_{48})$
M	x_{69}	y_{69}
N	$\frac{x_{69}+x_{70}}{2}$	$\frac{y_{69}+y_{70}}{2}$
O	$\frac{x_{70}+x_{71}}{2}$	$\frac{y_{70}+y_{71}}{2}$
P	x_{72}	y_{72}
Q	x_{73}	y_{73}
R	x_{74}	y_{74}
S	$\frac{x_{75}+x_{76}}{2}$	$\frac{y_{75}+y_{76}}{2}$
T	x_{77}	y_{77}
U	$\frac{x_{77}+x_{78}}{2}$	$\frac{y_{77}+y_{78}}{2}$
V	x_{79}	y_{79}
W	x_{80}	y_{80}
X	x_{65}	y_{65}
Y	$\frac{x_{68}+x_{67}}{2}$	$\frac{y_{68}+y_{67}}{2}$
Z	$\frac{x_{69}+x_{68}}{2}$	$\frac{y_{69}+y_{68}}{2}$

pois pertencem às íris e 26 pontos são obtidos pelas equações da Tabela 4.1.

Para que as máscaras $\mathcal{K}'$ sejam úteis para provimento de informações para animação de um avatar tridimensional, elas devem ser reescaladas e colocadas no mesmo sistema de referência da máscara de controle padrão da Figura 4.10. Para tanto, novas análises de Procrustes são realizadas, utilizando-se os cantos dos olhos (índices $k=21$ e $k=19$), ponta do nariz ($k=57$), narinas ($k=60$ e $k=59$) e ponto intermediário do lábio superior ($k=48$) para a obtenção dos parâmetros de transformação rígida. Obtém-se assim outras matrizes de rotação $\mathbf{R}_i$, outros vetores de translação $\mathbf{t}_i$ e escalares s_i (com $i=1,...,n$) e aplica-se a transformação da Equação 4.1 a cada conjunto de vértices da máscara $\mathcal{K}'$, obtendo-se conjuntos $\mathcal{K}$ na mesma escala e posição da máscara de controle padrão do sistema de animação. Para cada ponto do conjunto $\mathcal{K}'=\{(x'_k,y'_k,z'_k)\}$, com $k=1,...,86$

é aplicada a transformação:

$$(x_k, y_k, z_k) = s_i.(x'_k, y'_k, z'_k).\mathbf{R}_i + \mathbf{t}_i. \tag{4.2}$$

Assim, foram obtidas n máscaras de controle $\mathcal{K}$, registradas com a máscara de controle de animação padrão, cada uma classificada de acordo com o FACS e com a expressão realizada pelos indivíduos do banco de dados Bosphorus. Essas máscaras serão utilizadas para obtenção de características de entrada para treinamento dos classificadores e para estimativa do deslocamento tridimensional dos pontos de controle na animação, conforme as Seções 4.1.5 e 4.2.3. Com essas n máscaras, conclui-se a etapa 6 da Figura 4.1.

4.1.5 Obtenção de Características Utilizadas como Entrada para Classificadores

A ideia central para a construção da estrutura de dados Persona e sua posterior utilização está fundamentada na classificação automática da unidade de ação ou emoção realizada pelo ator A e pelo usuário U, para que a unidade de ação ou emoção realizada por U seja manifestada no avatar, usando máscaras que representam o estilo de movimento de A.

Logo, é essencial propor uma estratégia para classificação automática das unidades de ação correspondentes aos pontos característicos rastreados pelo *Live Driver*. Assim, quando um novo conjunto de pontos rastreados de um ator A (Λ_A) ou de um usuário U (Λ_U) for obtido, deve-se automaticamente determinar as unidades de ação ou emoção associadas. As Figuras 3.20 a 3.22 apresentam imagens com a descrição das unidades de ação e emoções do FACS anotadas no banco de dados Bosphorus e que são utilizadas nesse trabalho.

Para tanto, decidiu-se realizar a caracterização das unidades de ação associadas a diferentes componentes da face independentemente: olho direito, olho esquerdo, sobrancelha direita, sobrancelha esquerda e boca. Essa decisão respalda-se no fato de que um usuário pode realizar combinações de unidades de ação diferentes das combinações que o ator realizou durante a construção da Persona. Por exemplo, ele pode piscar com um olho ($UFAU\_43$), manter o outro aberto (expressão neutra), manter as sobrancelhas em repouso (expressão neutra) e sorrir (E_HAPPY). Tal combinação pode não ter sido realizada pelo ator durante a construção da Persona, mas dever ser reproduzida no avatar. Assim, o conjunto Λ foi subdividido em Λ_B (pontos pertencentes à boca), Λ_{Oe} (pontos do olho esquerdo), Λ_{Od} (pontos do olho direito), Λ_{Se} (pontos da sobrancelha esquerda) e Λ_{Sd} (pontos da sobrancelha direita). Esses subconjuntos de pontos podem ser visualizados na Figura 4.12. Não será considerado o movimento dos pontos pertencentes ao nariz, pois esse possui menor flexibilidade de movimentação. Essa divisão da face será utilizada tanto no processo de treinamento supervisionado dos classificadores, que é baseado na anotação do banco de dados Bosphorus, como na classificação das unidades de ação dado um novo ator ou usuário.

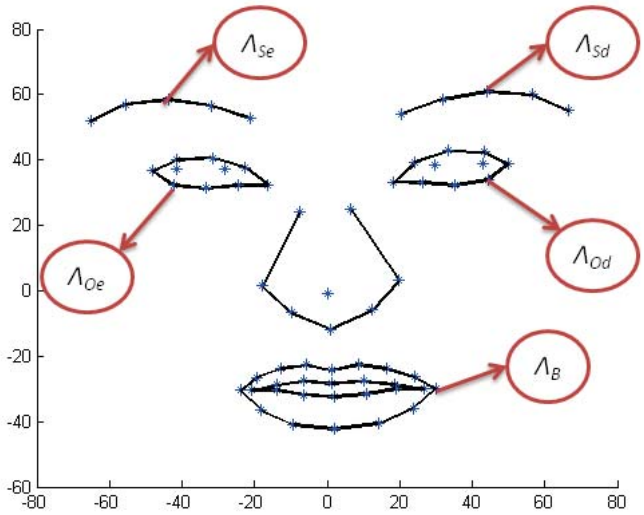


Figura 4.12: Divisão dos pontos do conjunto Λ em componentes faciais.

Obtenção de Descritores das Unidades de Ação ou Emoção

As diferentes características morfológicas dos componentes faciais tornam difícil a discriminação de unidades de ação considerando apenas as posições absolutas dos pontos. Isso decorre da variabilidade de fatores como espessura dos lábios, largura da boca, tamanho dos olhos, mobilidade das sobrancelhas e proximidade da face à câmera no momento da aquisição das imagens. Assim, verificou-se que as unidades de ação são melhor discrimináveis se considerarmos os **deslocamentos** dos pontos de controle em relação à expressão neutra. Para isso, é importante que tais deslocamentos estejam na mesma escala e posição para que se tornem comparáveis. Dado que foi feita a análise de Procrustes em todas as máscaras, assume-se que eles estão em mesma escala. Quanto à posição, cada subconjunto é deslocado para pontos específicos conforme descrição a seguir.

Para cada pessoa P cuja nuvem de pontos da face foi fornecida no Bosphorus, é identificada a máscara correspondente à expressão neutra $\mathcal{K}_{NP}$. As máscaras das demais expressões da pessoa P serão designadas por $\mathcal{K}_P$. Consideremos inicialmente o treinamento das unidades de ação e emoções da boca. Nesse caso, somente o subconjunto de pontos pertencentes à boca $\mathcal{K}_{P_B}$ será considerado. Tais pontos de todas as máscaras da pessoa P são então deslocados, de forma que todos fiquem com o ponto 48 da boca (centro do lábio superior) no mesmo local, conforme mostra a imagem superior da Figura 4.13. Esse ponto será denominado **ponto âncora** da boca e corresponde ao ponto $k = 48$ da máscara de controle padrão.

Da mesma forma que para a componente facial boca, os pontos dos olhos e sobrancelhas esquerdas são deslocados de forma que os cantos internos dos olhos de todas as máscaras estejam no mesmo local. Assim, para o olho e sobrancelha esquerdos o ponto âncora será o canto interno do olho esquerdo da máscara de controle e para o olho e sobrancelha direitos o ponto âncora será o canto interno do olho direito da máscara de controle padrão. As imagens inferiores da Figura 4.13

mostram os pontos deslocados, respectivamente, dos olhos e sobrancelhas esquerdas e dos olhos e sobrancelhas direitas.

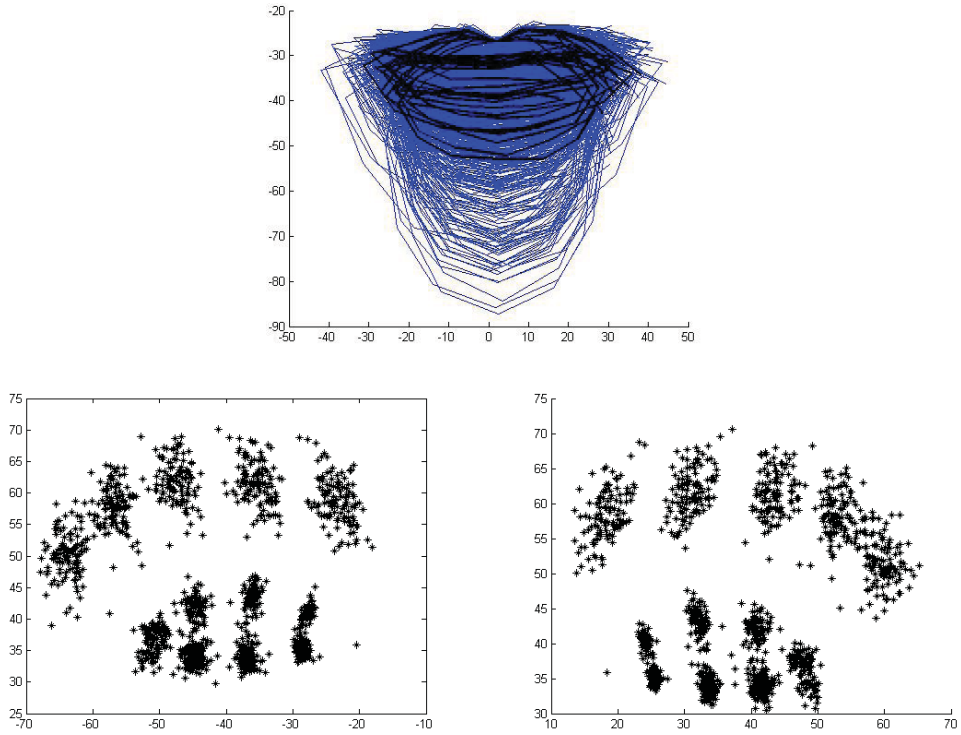


Figura 4.13: Acima, coordenadas horizontal e vertical dos pontos das máscaras de controle pertencentes à boca, com ponto âncora no vértice número 48 da máscara de controle padrão. Abaixo e à esquerda, coordenadas horizontal e vertical dos pontos dos olhos e sobrancelhas esquerda, com ponto âncora no canto interno do olho esquerdo. Da mesma forma, para os pontos dos olhos e sobrancelhas direitas, na imagem da direita.

Para cada i-ésimo conjunto de vértices da máscara de controle ($\mathcal{K}_{P_i}$) das expressões digitalizadas da pessoa P, são calculados os deslocamentos dos vértices em relação à expressão neutra $\mathcal{K}_{NP}$, conforme o conjunto de Equações 4.3. Nessas equações, são determinadas cinco matrizes de deslocamentos: $\mathbf{\Delta_{B_i}}$ para os deslocamentos verticais e horizontais dos pontos característicos da boca, $\mathbf{\Delta_{Oe_i}}$ para os deslocamentos verticais e horizontais dos pontos característicos do olho esquerdo e $\mathbf{\Delta_{Od_i}}$ para os do olho direito, $\mathbf{\Delta_{Se_i}}$ para os deslocamentos verticais e horizontais dos pontos característicos da sobrancelha esquerda e $\mathbf{\Delta_{Sd_i}}$ para os da sobrancelha direita.

$$
\begin{aligned}
\mathbf{\Delta_{B_i}} &= \mathcal{K}_{P_{B_i}} - \mathcal{K}_{NP_B}, \\
\mathbf{\Delta_{Oe_i}} &= \mathcal{K}_{P_{Oe_i}} - \mathcal{K}_{NP_{Oe}}, \\
\mathbf{\Delta_{Od_i}} &= \mathcal{K}_{P_{Od_i}} - \mathcal{K}_{NP_{Od}}, \\
\mathbf{\Delta_{Se_i}} &= \mathcal{K}_{P_{Se_i}} - \mathcal{K}_{NP_{Se}}, \\
\mathbf{\Delta_{Sd_i}} &= \mathcal{K}_{P_{Sd_i}} - \mathcal{K}_{NP_{Sd}}.
\end{aligned}
\tag{4.3}
$$

Nessas equações, $\mathcal{K}_{\mathbf{NP_c}}$ e $\mathcal{K}_{\mathbf{P_c}}$ representam, respectivamente, os pontos dos componentes faciais da máscara neutra e das máscaras não neutras da pessoa P, para $c \in \{B, Oe, Od, Se, Sd\}$. O índice i varia de 1 até o número de expressões disponíveis da pessoa P. A letra B indica que são considerados apenas os pontos da boca, Oe do olho esquerdo, Od do olho direito, Se da sobrancelha esquerda e Sd da sobrancelha direita.

Após obtidas as matrizes de deslocamentos horizontal e vertical ($\mathbf{\Delta}s$) das Equações 4.3, as duas primeiras colunas dessas matrizes são concatenadas. Com esse procedimento, são obtidos n conjuntos de cinco vetores de deslocamentos horizontais e verticais. Esses cinco vetores correspondem aos cinco componentes faciais e serão designados por $\boldsymbol{\delta'}_{B_j}$, $\boldsymbol{\delta'}_{Oe_j}$, $\boldsymbol{\delta'}_{Od_j}$ $\boldsymbol{\delta'}_{Se_j}$ e $\boldsymbol{\delta'}_{Sd_j}$, com $j = 1, ..., n$. Os deslocamentos em profundidade são desconsiderados para o treinamento dos classificadores, pois não haverá disponibilidade dessas informações quando eles forem utilizados para classificar expressões provenientes de vídeo.

Cada elemento e_k de cada vetor $\boldsymbol{\delta'}$ deve então ser padronizado. O processo de padronização[3] permite comparar dados provenientes de distribuições distintas, visto que leva em conta o desvio padrão de cada elemento do vetor. Assim, deslocamentos de pontos com maior mobilidade podem ser comparados com deslocamentos de pontos anatomicamente com menor mobilidade. Para essa etapa de padronização, calcula-se a média dos elementos dos n vetores de cada componente facial ($\bar{\boldsymbol{\delta}}_B$,$\bar{\boldsymbol{\delta}}_{Oe}$, $\bar{\boldsymbol{\delta}}_{Od}$, $\bar{\boldsymbol{\delta}}_{Se}$ e $\bar{\boldsymbol{\delta}}_{Sd}$) e os respectivos desvios padrão ($\boldsymbol{\sigma}_B$,$\boldsymbol{\sigma}_{Oe}$, $\boldsymbol{\sigma}_{Od}$, $\boldsymbol{\sigma}_{Se}$ e $\boldsymbol{\sigma}_{Sd}$). Aplica-se, então, a equação de padronização para cada elemento e_k de cada j-ésimo vetor $\boldsymbol{\delta'}$:

$$
\begin{aligned}
\delta_{B_{jk}} &= \frac{e_{B_{jk}} - \bar{e}_{B_k}}{\sigma_{B_k}}, \\
\delta_{Oe_{jk}} &= \frac{e_{B_{jk}} - \bar{e}_{Oe_k}}{\sigma_{Oe_k}}, \\
\delta_{Od_{jk}} &= \frac{e_{B_{jk}} - \bar{e}_{Od_k}}{\sigma_{Od_k}}, \\
\delta_{Se_{jk}} &= \frac{e_{B_{jk}} - \bar{e}_{Se_k}}{\sigma_{Se_k}}, \\
\delta_{Sd_{jk}} &= \frac{e_{B_{jk}} - \bar{e}_{Sd_k}}{\sigma_{Sd_k}}.
\end{aligned}
\tag{4.4}
$$

Assim, obtém-se, para cada componente facial, n vetores de deslocamento padronizados, designados por $\boldsymbol{\delta_B}$, $\boldsymbol{\delta_{Oe}}$, $\boldsymbol{\delta_{Od}}$, $\boldsymbol{\delta_{Se}}$ e $\boldsymbol{\delta_{Sd}}$. Esses vetores serão submetidos à análise de componentes principais, conforme explicado a seguir.

Análise de Componentes Principais

Uma vez obtidos esses vetores para todas as n máscaras de controle, é realizada a análise de componentes principais. Para cada componente facial é realizada uma PCA. Dessa forma, cada

[3] Também conhecido como cálculo do escore Z, onde a diferença entre cada dado x e a média μ é dividida pelo desvio padrão σ ($Z = \frac{x-\mu}{\sigma}$).

expressão da boca, por exemplo, pode ser representada por um ponto no subespaço dos componentes principais, conforme mostra a Figura 4.14.

Apenas a título de visualização, membros de algumas classes (correspondentes à anotação do Bosphorus) foram coloridos em amarelo (LFAU_12), preto (E_FEAR), verde (E_HAPPY), vermelho (LFAU_10), ciano (E_SURPRISE) e magenta (E_SADNESS). Pode-se observar na Figura 4.14 que existem agrupamentos de representantes de mesma classe. Por exemplo, os deslocamentos da boca correspondentes à classe E_HAPPY (em verde) tendem a ter um valor do segundo componente principal relativamente elevado quando comparados com as demais classes (E_SADNESS, por exemplo). Pode-se verificar também, entretanto, que as classes não são linearmente separáveis, ou seja, há sobreposição entre os agrupamentos. Essa questão será detalhadamente discutida mais adiante.

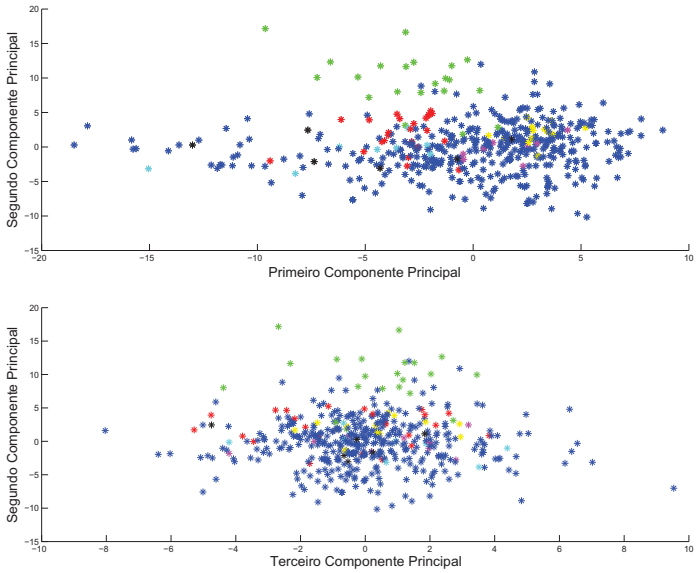


Figura 4.14: Visualização das expressões da boca do banco de dados Bosphorus expressas em termos dos três componentes principais.

Para ilustrar o efeito da análise componente principal, a Figura 4.15 mostra a variação da classe "Emotion Happy" (E_HAPPY) ao longo dos eixos dos dois componentes principais. Podem-se observar, por exemplo, sorrisos com lábios mais fechados à esquerda (menores coeficientes do primeiro PC) e sorrisos com lábios mais abertos à direita (maiores coeficientes do primeiro PC).

A fim de determinar o número de dimensões do subespaço de componentes principais que representa uma parcela significativa da variabilidade dos dados, verificou-se a variabilidade acumulada representada pelos PC. A Figura 4.16 mostra o percentual da variabilidade acumulada em função do número de componentes principais utilizados para compor o subespaço. Tomou-se como exemplo o caso do componente facial boca. Conforme mostra essa figura, se o subespaço de componentes principais tiver uma única dimensão, apenas 38,2% da variabilidade das expressões da boca serão representadas; se dois componentes principais forem considerados, teremos 68,3% da variabilidade dos dados representada. Note que, ao levarmos em consideração mais componentes principais, acrésci-

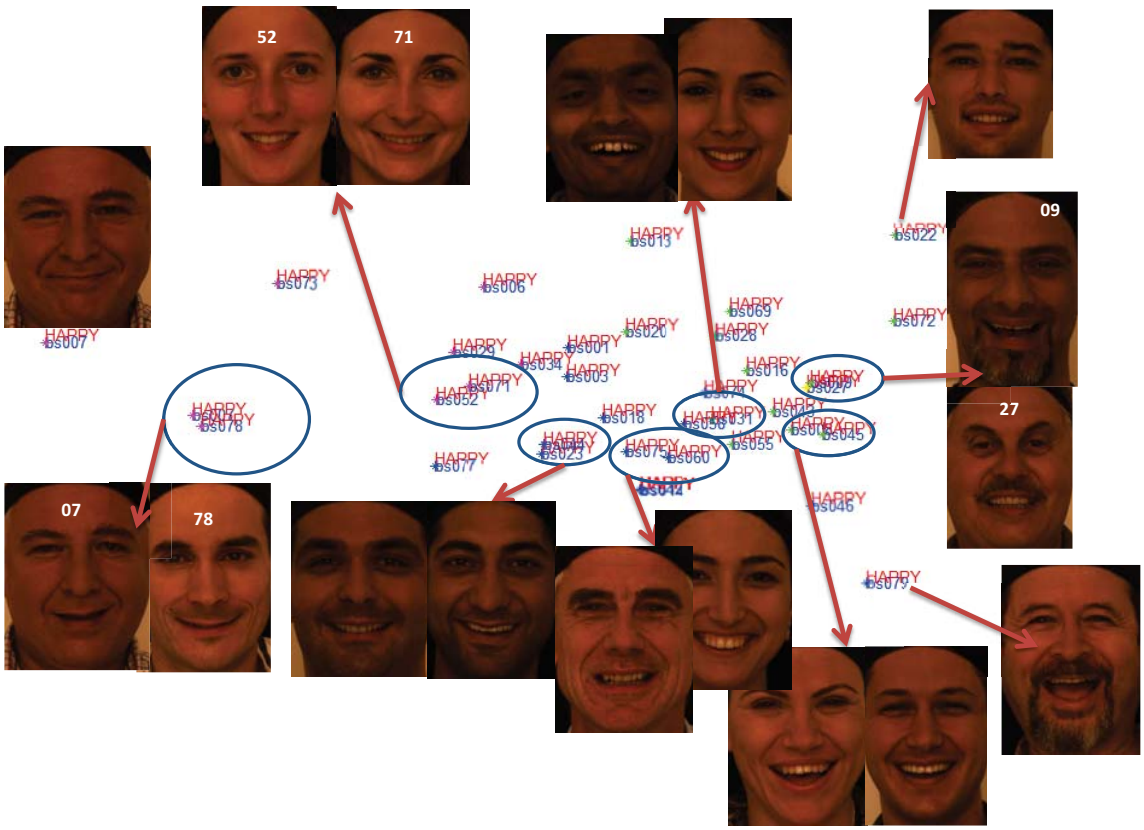


Figura 4.15: Visualização do efeito da variação dos coeficientes dos dois primeiros componentes principais, considerando apenas a emoção alegria.

mos cada vez menores no percentual de variabilidade dos dados vão sendo acumulados. Assim sendo, decidiu-se que o subespaço de componentes principais terá dimensão $D = 11$, que corresponde a 93,78% da variabilidade dos dados. Dessa forma, cada unidade de ação ou emoção da boca $\mathcal{E}_j$ fica

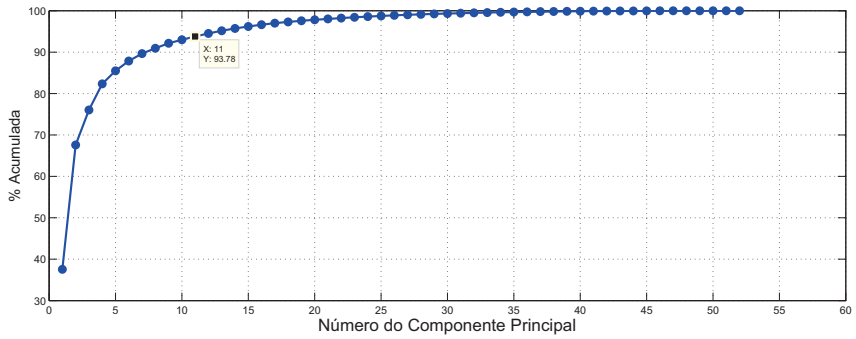


Figura 4.16: Porcentagem acumulada da variabilidade dos dados referentes ao movimento da boca. Note que os 11 primeiros componentes principais acumulam mais de 93% da variabilidade dos dados.

representada por um ponto no espaço 11-dimensional dado pela expressão:

$$\mathcal{E}_j = \sum_{d=1}^{11} a_d . PC_d, \tag{4.5}$$

em que a_d são os coeficientes de cada PC. O vetor de coeficientes $\mathcal{A}_\mathbf{j} = \{a_1, a_2, ..., a_{11}\}$ passa então a caracterizar cada expressão da boca. Análise semelhante foi feita para os demais componentes faciais (olhos e sobrancelhas) e decidiu-se que o subespaço principal para olhos e sobrancelhas terá dimensão $D = 5$, o que corresponde a mais de 95% da variabilidade dos dados. Esses vetores são armazenados para posterior utilização, sendo uma das saídas da fase de pré-processamento, conforme Figura 4.1. Com isso, encerra-se o processo de número 7 representado nessa figura.

Embora a análise de componentes principais tenha se mostrado útil para a classificação das unidades de ação ou emoção de cada componente facial, notou-se que uma simples medida de distância no subespaço principal (mesmo uma distância que leve em conta a relevância de cada componente principal na variabilidade dos dados) não foi suficiente para classificar satisfatoriamente novas expressões. Isso ocorre porque os agrupamentos de expressões pertencentes à mesma classe não são linearmente separáveis, conforme observado na Figura 4.14. Por isso, técnicas como *k-means* [42] utilizando distâncias Euclidiana e de Mahalanobis [43] foram testadas mas não se mostraram eficientes na classificação. A Figura 4.14 mostra os escores de 525 expressões de boca nos dois primeiros componentes principais na imagem superior e do segundo e terceiro componente principal na imagem inferior. Apesar de apenas 3 componentes principais serem mostrados (o que representa em torno de 76% da variabilidade dos dados) pode-se observar que as classes não são linearmente separáveis, pois há sobreposição de agrupamentos de classes distintas.

As imagens da Figura 4.17 mostram as quatro expressões do banco de dados Bosphorus com pontos no subespaço principal mais próximos ao ponto correspondente à unidade de ação ou emoção dos lábios realizada pela pessoa da esquerda. Em outras palavras, são mostrados os 4 vizinhos mais próximos da representação da expressão dos lábios do usuário em questão no subespaço principal. Foi utilizada a distância Euclidiana nesse subespaço para obtenção desses vizinhos. Essa figura está organizada da seguinte forma: os usuários são mostrados nas imagens maiores à esquerda e as 4 expressões mais próximas são mostradas na ordem esquerda superior, direita superior, esquerda inferior e direita inferior.

Pode-se considerar que as quatro expressões do Bosphorus mais próximas ao sorriso do usuário da imagem A da Figura 4.17 no subespaço principal correspondem satisfatoriamente à ação da boca do usuário. Entretanto, a inspeção visual das imagens B, C e D mostra que pontos próximos no subespaço principal nem sempre pertencem à mesma classe da ação que o usuário está realizando com a boca. No caso da imagem B, a segunda e terceira expressões mais próximas não são correspondentes à unidade de ação executada pela usuária. No caso da imagem C, a quarta expressão mais próxima à do usuário também não é adequada. Já para a imagem D, observa-se que a segunda e quarta expressões mais próximas à unidade de ação do usuário estão mais adequadas do que a primeira. Isso se deve ao fato de que as classes não são linearmente separáveis e de que, às vezes, características representadas em um dado componente principal são mais relevantes para a classificação de uma classe específica do que as demais.

Assim sendo, optou-se por utilizar outra ferramenta de classificação capaz de diferenciar classes não linearmente separáveis no espaço dos componentes principais. A ferramenta escolhida dado

Figura 4.17: Imagens do banco de dados Bosphorus cujas expressões da boca são mais próximas à expressão do usuário no subespaço principal.

o histórico de sucesso em trabalhos de reconhecimento de unidades de ação e expressões faciais (conforme mostrado na Seção 2.2) foram as redes neurais artificiais. Esse procedimento será descrito

a seguir.

4.1.6 Redes Neurais Artificiais para Reconhecimento de Unidades de Ação ou Expressões

Dado que cada conjunto de pontos característicos de cada componente facial (boca, olho esquerdo, olho direito, sobrancelha esquerda e sobrancelha direita) foi transformado em um vetor de coeficientes $\mathcal{A} = \{a_d\}$, com $d = 1, 2, ..., D$, os quais representam cada expressão no subespaço D-dimensional de componentes principais, pode-se utilizar esses vetores como características de entrada em classificadores automáticos.

Conforme Seção 2.2, há um histórico de sucesso na literatura da aplicação de redes neurais para reconhecimento de unidades de ação da face. Logo, decidiu-se utilizar redes neurais artificiais como classificadores automáticos das unidades de ação ou emoção. Nos testes realizados, foram utilizados para treinamento os vetores $\mathcal{A}$ com os coeficientes dos 11 primeiros componentes principais para a boca e 5 componentes principais para as sobrancelhas esquerda e direita e para os olhos direito e esquerdo, conforme seção anterior.

Verificou-se empiricamente que uma rede neural *feedforward*, com algoritmo de treinamento *backpropagation*, apresentou desempenho satisfatório para os propósitos desse trabalho. Foi utilizada apenas uma camada oculta, com 30 neurônios para o caso da boca e 20 neurônios para os demais componentes faciais. Mais camadas e/ou neurônios tendem a fazer com que a rede se adapte em demasia aos dados de treinamento, perdendo sua capacidade de generalização conforme explicado na Seção 3.3.2. Essa característica é conhecida como *overtrainning* [4]. RNAs com menos neurônios na camada oculta tendem a não ter bom desempenho no processo de classificação.

Durante a fase de aprendizado das redes, foram treinadas várias RNAs com diferentes inicializações, até que condições de parada estabelecidas fossem atingidas. Tais condições foram de que a rede tivesse mais de 60% de acerto para a boca e mais de 90% de acerto para olhos e sobrancelhas. Essas redes foram então salvas para posterior utilização, constituindo-se em uma das saídas da etapa de pré-processamento ilustradas na Figura 4.1. A diferença no limiar de aceitação para o desempenho das RNAs da boca em relação às demais deve-se à maior probabilidade de confusão na classificação desse componente facial devido ao grande número de AUs ou emoções e ambiguidade na classificação. Exemplos dessas ambiguidades podem ser vistos na Figura 4.18.

Nos exemplos ilustrados na Figura 4.18 são mostradas classes com alta probabilidade de confusão entre si. Como são considerados apenas os deslocamentos dos pontos do contorno dos lábios para a classificação, a boca semi-aberta do caso A (AU $LFAU\_26$) pode ser confundida com a emoção surpresa ($E\_SURPRISE$). Nos casos B e C, dois indivíduos distintos do banco de dados Bosphorus estão executando as unidades de ação $LFAU\_34$ e $LFAU\_23$. A diferença entre as duas unidades de ação está no fato de que, na $LFAU\_34$, a compressão dos lábios é acompanhada do movimento de inflar a bochecha. Como os pontos da bochecha não são levados em conta nos classificadores, há alta probabilidade de confusão entre as duas classes. Nesses casos, pode-se ponderar que nem mesmo uma pessoa poderia diferenciar as classes da coluna da esquerda das classes da coluna da direita, se forem considerados apenas o contorno dos lábios para a análise. Assim sendo, decidiu-se

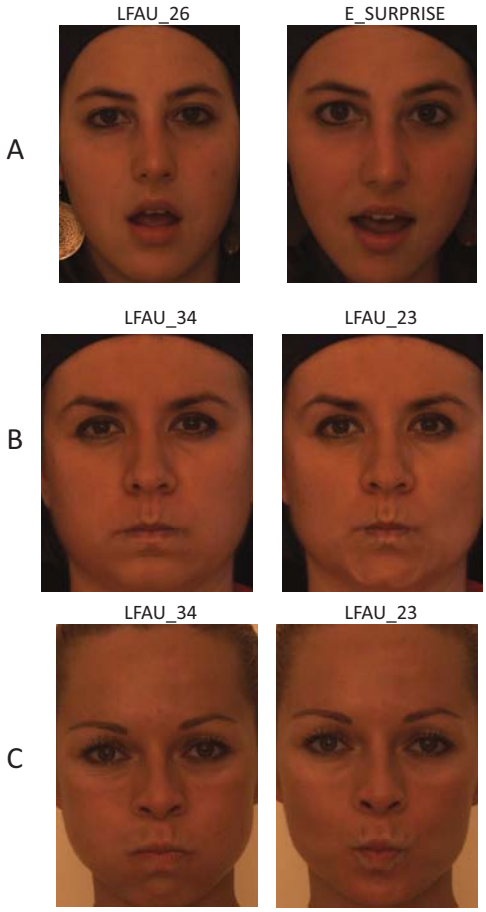


Figura 4.18: Exemplos de classes em que é provável confusão na classificação por parte das RNAs.

que ambas as classes são aceitáveis.

Tendo sido treinados os classificadores, chega-se ao final da fase de pré-processamento. Segue-se o processo de construção da Persona de atores, que será detalhado na próxima seção.

4.2 Construção da Persona

Para a construção da Persona de um ator A, devem-se seguir quatro etapas que serão descritas a seguir: obtenção do vetor de *entradas das RNAs*, *classificação* das unidades de ação ou emoção, obtenção de *máscaras* tridimensionais associadas e *construção* da Persona propriamente dita.

Primeiramente, selecionam-se trechos de vídeo em que o ator A aparece com a face paralela ao plano da imagem, ou com pequeno ângulo de rotação. Aplica-se então a ferramenta *Live Driver* nessas sequências de vídeo, obtendo-se Q conjuntos de pontos característicos Λ_I, um para cada quadro q rastreado. O índice I indica que os pontos desses conjuntos estão em coordenadas de imagem. Dados esses Q conjuntos, escolhe-se manualmente um referente a um quadro em que o ator esteja com expressão neutra. O conjunto Λ_I referente a esse quadro será denominado Λ_{IN}. Após, prossegue-se na execução das quatro etapas listadas anteriormente, conforme as próximas Seções (4.2.1 a 4.2.4).

4.2.1 Obtenção do Vetor de Entradas das RNAS

Da mesma forma realizada com os pontos de controle rastreados no banco de dados Bosphorus, é feito o registro dos pontos rastreados de acordo com a máscara padrão da Figura 4.10. Para tanto, é realizada a análise de Procrustes, explicada na Seção 3.6, para escalar e rotacionar os pontos de cada conjunto Λ_{Iq} de acordo com a máscara de controle padrão, obtendo-se os conjuntos Λ_q (com $q = 1, ..., Q$) de 64 pontos. Em seguida, são separados os pontos de cada um dos Q conjuntos Λ que pertencem à boca (Λ_B), ao olho esquerdo (Λ_{Oe}) ao olho direito (Λ_{Od}), à sobrancelha esquerda (Λ_{Se}) e à sobrancelha direita (Λ_{Sd}), conforme a Figura 4.12.

A partir desses subconjuntos, para cada componente facial, o procedimento segue de acordo com o ilustrado na Figura 4.19, a fim de que sejam obtidos os vetores de características que serão entrada da rede neural artificial. Essa figura utiliza como exemplo a análise dos pontos pertencentes à boca.

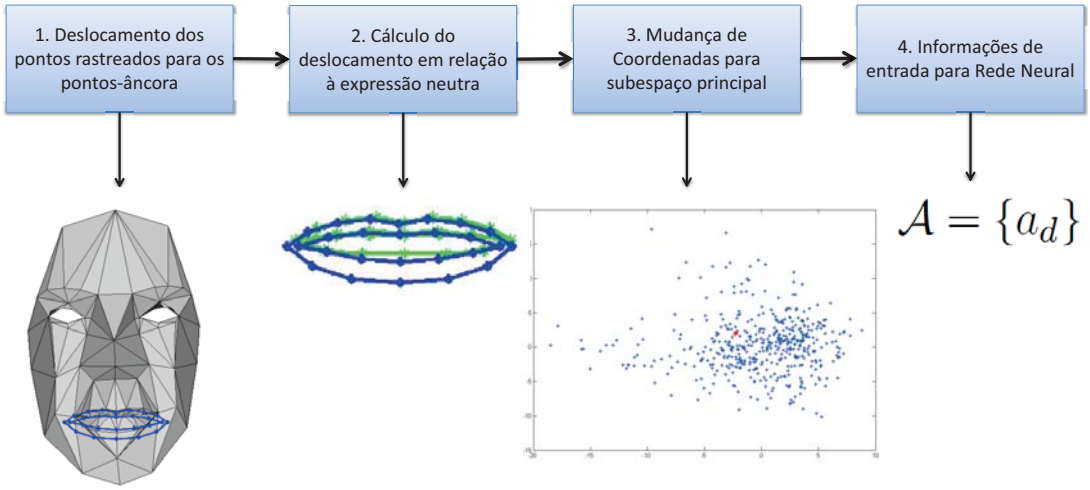


Figura 4.19: Sequência de passos para obtenção do vetor de características de entrada das RNAs. Essa figura considera apenas o componente facial boca. Processo semelhante ocorre para os demais componentes faciais.

De acordo com o esquema da Figura 4.19, o processo de obtenção dos vetores de características que serão dados de entrada para as redes neurais começa com o deslocamento dos conjuntos de pontos de cada componente facial para os respectivos pontos âncora. Relembrando, para a boca o ponto âncora é o ponto intermediário do lábio superior. Para os olhos e sobrancelhas, os pontos âncoras são os cantos internos dos olhos. Dessa forma, serão sobrepostos os pontos pertencentes à expressão neutra do ator Λ_N e os pontos da expressão rastreada no q-ésimo quadro da sequência de vídeo Λ_q. Na Figura 4.19 é ilustrado o deslocamento dos pontos do conjunto Λ_{B_q} para o ponto âncora da boca da máscara de controle (ponto intermediário do lábio superior). Da mesma forma, os pontos rastreados da expressão neutra referentes à boca (Λ_{NB}) são deslocados para o mesmo ponto âncora. Os pontos referentes à boca do conjunto Λ_{NB} são mostrados conectados em verde

na imagem da segunda etapa dessa figura.

Em seguida, ainda de acordo com a Figura 4.19, são calculados os deslocamentos horizontal e vertical dos pontos de cada componente facial da q-ésima expressão do ator A em relação aos pontos de cada componente facial da face neutra, com o processo descrito no conjunto de Equações 4.3. As colunas das matrizes de deslocamento são então concatenadas e os vetores assim obtidos são padronizados, utilizando-se médias ($\bar{\boldsymbol{\delta}}_B$, $\bar{\boldsymbol{\delta}}_{Od}$, $\bar{\boldsymbol{\delta}}_{Oe}$, $\bar{\boldsymbol{\delta}}_{Sd}$ e $\bar{\boldsymbol{\delta}}_{Se}$) e desvios padrão ($\boldsymbol{\sigma}_B$, $\boldsymbol{\sigma}_{Od}$, $\boldsymbol{\sigma}_{Oe}$, $\boldsymbol{\sigma}_{Sd}$ e $\boldsymbol{\sigma}_{Se}$) obtidos durante o processo de treinamento. Esse processo de padronização é idêntico ao descrito pela Equação 4.4.

Esses vetores devem, então, ser aproximados por um ponto no subespaço dos componentes principais obtidos também na fase de treinamento. Para isso, deve-se lembrar de que os componentes principais obtidos são base do subespaço principal. Nesse procedimento, calcula-se primeiramente a matriz $\mathbf{M}$ de mudança de base da base canônica do $\Re^d$ ($\mathbf{E}$) para a base $\mathbf{P}$ dos D componentes principais estabelecidos. Essa matriz é obtida pela resolução da equação:

$$\mathbf{E}.\mathbf{M} = \mathbf{P}. \tag{4.6}$$

Essa equação pode ser resolvida por fatorização LU [72]. De posse da matriz de mudança de base, podem-se calcular o vetor de coeficientes $\mathcal{A} = \{a_d\}$ referentes à expressão adotada pelo ator A em cada componente facial no subespaço principal pelas equações:

$$\begin{aligned}
\mathcal{A}_{B_q} &= \mathbf{M}_B.\boldsymbol{\delta}_{B_q}, \\
\mathcal{A}_{Oe_q} &= \mathbf{M}_{Oe}.\boldsymbol{\delta}_{Oe_q}, \\
\mathcal{A}_{Od_q} &= \mathbf{M}_{Od}.\boldsymbol{\delta}_{Od_q}, \\
\mathcal{A}_{Se_q} &= \mathbf{M}_{Se}.\boldsymbol{\delta}_{Se_q}, \\
\mathcal{A}_{Sd_q} &= \mathbf{M}_{Sd}.\boldsymbol{\delta}_{Sd_q}.
\end{aligned} \tag{4.7}$$

Na Figura 4.19, esse processo é exemplificado para o componente facial boca. O ponto em vermelho no gráfico da terceira etapa dessa figura corresponde aos coeficientes dos dois primeiros componentes principais, que representam a unidade de ação ou emoção da boca no quadro q. Os demais pontos mostrados nesse gráfico em azul são correspondentes às unidades de ação ou emoção provenientes do Bosphorus. Logo, a unidade de ação ou emoção caracterizada pelo resultado do rastreamento do quadro q mostrado em azul na imagem da segunda etapa da Figura 4.19 fica representada pelo vetor de coeficiente de $D = 11$ dimensões $\mathcal{A}_{B_q}$.

4.2.2 Classificação das Unidades de Ação ou Emoção

Uma vez obtidos os vetores de características $\mathcal{A}_\mathbf{q} = \{a_d\}$ (com $d = 1, ..., D$) de cada componente facial de acordo com o conjunto de Equações 4.7 no quadro q e tendo-se treinado as 5 redes neurais (uma para cada componente facial de interesse), prossegue-se para a classificação desses vetores pelas RNAs. Um esquema para facilitar a compreensão da interpretação da saída das redes neurais

pode ser visualizado da Figura 4.20. Cada RNA retorna como saída um vetor $\boldsymbol{\varphi} = \boldsymbol{F}(\boldsymbol{\rho})$ com valores no intervalo [-1,1], cuja dimensão corresponde ao número de classes (AUs ou Emoções anotados no Bosphorus) passíveis de serem identificadas de acordo com os dados de treinamento. Nessa figura, $c_1, c_2, ..., c_n$ são os nomes das classes consideradas durante o treinamento, que, em suma, correspondem às AUs ou emoções do banco de dados Bosphorus, anotadas de acordo com o FACS. $\varphi_1, \varphi_2, ..., \varphi_n$, são os valores do intervalo [-1,1] retornado pelas redes neurais.

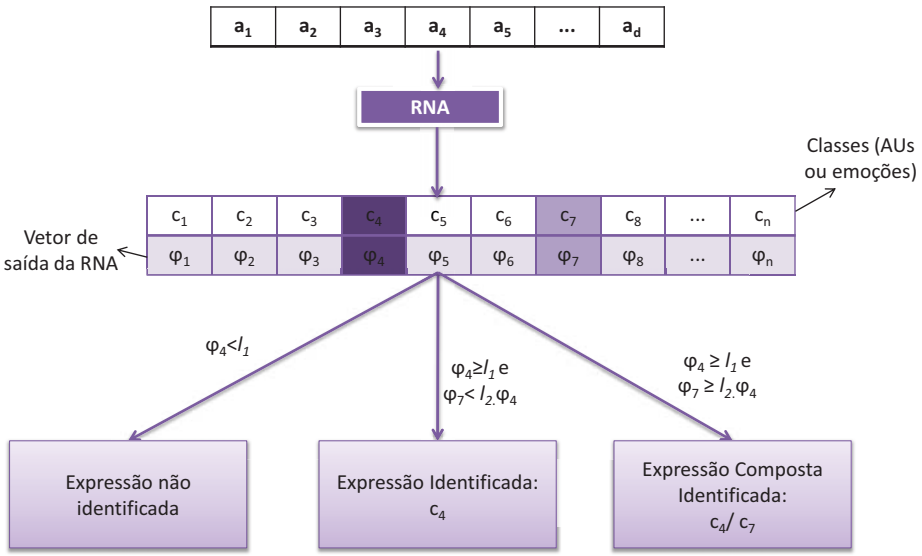


Figura 4.20: Interpretação do resultado da RNA e atribuição da classificação à expressão rastreada.

A classe c_i associada ao maior valor do vetor $\boldsymbol{\varphi}$ indica qual a unidade de ação foi classificada como sendo a ação executada pelo usuário U. Essa classe será denominada "Expressão Principal" e será designada por c_{p_1}. Da mesma forma, o segundo maior valor do vetor $\boldsymbol{\varphi}$ pode também indicar a presença da respectiva unidade de ação ou emoção na expressão que o ator está executando nesse quadro. A classe correspondente a esse segundo maior valor será considerada a "Expressão Secundária", designada por c_{p_2}. A interpretação dos resultados das RNAs é feita como segue:

- Se o valor atribuído pela RNA na posição c_{p_1} do vetor $\boldsymbol{\varphi}$, $\varphi_{c_{p_1}}$, for menor que um limiar arbitrário $l_1 \in [-1, 1]$, não é considerado sucesso na classificação e a expressão desse quadro é considerada desconhecida.

- Caso seja identificado que c_{p_1} é uma classe conhecida, verifica-se se $\varphi_{c_{p_2}}$ é maior que $l_2.\varphi_{c_{p_1}}$, com $l_2 \in [0, 1]$ sendo um segundo limiar arbitrário. Caso isso ocorra, a expressão é considerada "Composta" e ambas são consideradas como unidades de ação ou emoção presentes na expressão rastreada. Caso contrário ($\varphi_{c_{p_2}} \leq l_2.\varphi_{c_{p_1}}$), apenas a unidade de ação ou emoção c_{p_1} será atribuída à expressão facial rastreada.

A Figura 4.20 ilustra o processo descrito nos itens acima, considerando a situação hipotética em que a classe principal corresponde à $c_{p_1} = c_4$ e a classe secundária corresponde à $c_{p_2} = c_7$. O

sucesso da classificação dependerá do valor de φ_4. Se esse valor for menor do que o limiar l_1, interpreta-se que a rede não foi capaz de classificar satisfatoriamente a unidade de ação ou emoção do ator *A*. Nesse caso, essa expressão não fará parte da Persona desse ator. Caso φ_4 seja maior ou igual a l_1, observa-se o valor de φ_7. Se ele for significativo quando comparado com o valor de $\varphi 4$ $(\varphi_7 > l_2.\varphi_4)$, diz-se que a expressão do ator será composta e ambas as classes c_4 e c_7 serão atribuídas à expressão do ator. Essa expressão será parte da Persona com a denominação c_4/c_7. Caso contrário $(\varphi_7 \leq l_2.\varphi_4)$, a única classe atribuída à expressão do ator será c_4. Nesse trabalho, foram atribuídos os valores $l_1 = 0,1$ e $l_2 = 0,6$. Esses valores foram considerados satisfatórios para os fins desse trabalho, após experimentos.

4.2.3 Construção de Máscaras Tridimensionais

De posse da classificação da expressão realizada pelo ator A, procede-se para a construção das máscaras de controle tridimensionais associadas aos pontos rastreados em cada quadro q. Para cada um dos cinco componentes faciais, uma máscara de controle é gerada partir dos 64 pontos bidimensionais do conjunto Λ_q. O procedimento de geração dessas máscaras com 86 vértices é um pouco diferenciado para classes com unidades de ação associadas à parte superior da face (U_FAUs) em relação às unidades de ação associadas à parte inferior da face (L_FAUs). A Tabela 4.2 apresenta resumidamente esse processo que será detalhadamente explicado na sequência.

Tabela 4.2: Resumo da atribuição de valores às máscaras de controle do ator A.

Tipo de Expressão	$\mathcal{K} \cap \Lambda$			$\mathcal{K} - \mathcal{K} \cap \Lambda$		
	x	y	z	x	y	z
U_FAUs	x_Λ	y_Λ	$z_{\mathcal{K}_0}$	Tabela 4.1	Tabela 4.1	$z_{\mathcal{K}_0}$
L_FAUs	x_Λ	y_Λ	$z_{\mathcal{K}_0} + \Delta z_{\mathcal{B}}$	Tabela 4.1	Tabela 4.1	$z_{\mathcal{K}_0} + \Delta z_{\mathcal{B}}$

Na Tabela 4.2, a interseção $\mathcal{K} \cap \Lambda$ representa os vértices da máscara que apresentam pontos correspondentes no conjunto Λ, enquanto $\mathcal{K} - \mathcal{K} \cap \Lambda$ representa o conjunto de vértices da máscara que não têm pontos associados no conjunto Λ. O símbolo $z_{\mathcal{K}_0}$ representa a coordenada de profundidade da máscara de controle padrão (designada por $\mathcal{K}_0$ e mostrada na Figura 4.10). Δz_B representa a variação da coordenada z entre uma máscara de uma pessoa P do Bosphorus com mesma expressão que o ator está realizando e a máscara neutra dessa mesma pessoa P. Finalmente, x_Λ e y_Λ são as coordenadas vertical e horizontal do conjunto Λ. Esses símbolos serão melhor definidos nos próximos parágrafos.

No caso das U_FAUs (unidades de ação dos olhos e sobrancelhas), os vértices das máscaras $\mathcal{K}_q$ que correspondem aos pontos do conjunto Λ_q recebem os valores das coordenadas horizontal x_{Λ_q} e vertical y_{Λ_q} dos pontos desse conjunto. As coordenadas x e y dos demais pontos da máscara são calculados conforme as equações da Tabela 4.1. A coordenada z recebe o mesmo valor da profundidade $z_{\mathcal{K}_0}$ dos vértices correspondentes da máscara padrão $\mathcal{K}_0$, visto que os deslocamentos no plano transverso ao plano da face das sobrancelhas e olhos são mínimos. Essas atribuições podem ser vistas resumidamente na linha correspondente ao tipo de expressão U_FAU da Tabela 4.2.

Da mesma forma, os vértices das máscaras correspondentes às L_FAUs, que possuem correspondência com os pontos do conjunto Λ_q assumem as coordenadas horizontal x_{Λ_q} e vertical y_{Λ_q} desses pontos. As coordenadas bidimensionais dos demais pontos são também determinadas de acordo com a Tabela 4.1. Resta, porém, conhecer os deslocamentos tridimensionais dos vértices associados à boca, queixo, bochechas e contorno da face (representados na Tabela 4.2 pelo subconjunto $(\mathcal{K} - \mathcal{K} \cap \Lambda)$). Esses deslocamentos tridimensionais são importantes pois são muito significativos para fins de animação da face. Como não há informação alguma da terceira coordenada desses componentes faciais no conjunto Λ_q, decidiu-se recorrer às informações do Bosphorus para inferir os deslocamentos humanamente aceitáveis nessa direção, para cada unidade de ação ou emoção cujas informações bidimensionais foram rastreadas. A hipótese aqui é de que ações semelhantes no plano da face são também semelhantes no plano perpendicular a ela.

Para isso, procede-se com a identificação de qual nuvem de pontos do Bosphorus pertencente à mesma classe principal identificada de acordo com a seção anterior tem vetores $\mathcal{A}_{B\mathcal{B}}$ mais próximos ao vetor de coeficientes da expressão em questão $\mathcal{A}_{Bk}$ no subespaço principal. $\mathcal{A}_{B\mathcal{B}}$ representa o vetor de coeficientes da boca (B) de uma dada pessoa do Bosphorus $(\mathcal{B})$. Essa nuvem de pontos corresponde a um indivíduo P do banco de dados, cuja máscara associada à expressão neutra também está disponível. Calcula-se, então, o deslocamento no eixo transverso ao plano da máscara da expressão neutra de P $(\mathcal{K}_{NP})$ para a máscara com expressão correspondente a $\mathcal{A}_{B\mathcal{B}}$, designada por $\mathcal{K}_P$. Esses deslocamentos, dados por $\mathbf{\Delta} z_{\mathcal{B}}$, são obtidos pelas subtrações:

$$\mathbf{\Delta} z_{\mathcal{B}} = \mathcal{K}_P - \mathcal{K}_{NP}. \tag{4.8}$$

Como todas as máscaras do Bosphorus estão na mesma escala da máscara padrão, os deslocamentos assim obtidos são aplicados diretamente na terceira coordenada dos pontos pertencentes aos lábios, queixo, contorno e bochechas da máscara padrão. Dessa forma são obtidos os vértices tridimensionais das cinco máscara de controle $\mathcal{K}_q$ referente ao q-ésimo quadro do vídeo do ator A.

A Figura 4.21 mostra uma máscara obtida por esse processo.

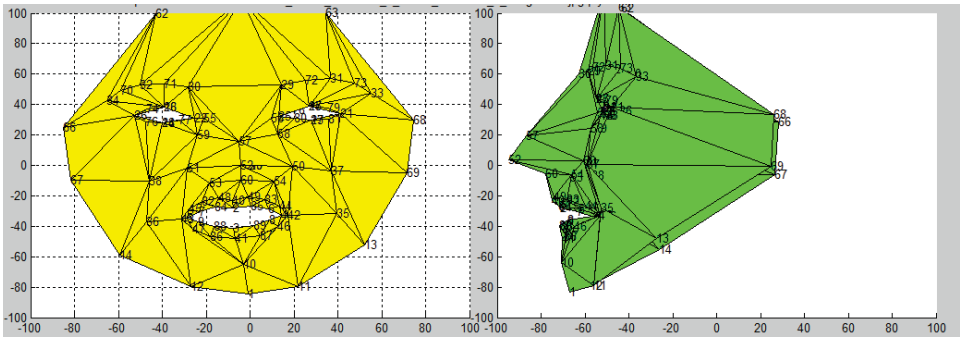


Figura 4.21: Máscara de controle construída a partir do rastreamento da face do ator e de informações de deslocamento em profundidade do Bosphorus. À esquerda, visão frontal da máscara associada à unidade de ação $LFAU\_22$. À direita sua visão lateral

4.2.4 A Estrutura Persona

As seções anteriores mostram como associar máscaras de controle de animação tridimensional à unidade de ação ou emoção realizada pelo ator. Todo esse processo é repetido para cada quadro do vídeo de entrada. À medida que as máscaras associadas a cada unidade de ação ou emoção vão sendo geradas, uma estrutura de dados conforme a Figura 4.22 vai sendo constituída. Nessa figura, cada célula representada por uma elipse corresponde a uma unidade de ação ou emoção classificada pelas RNAs. Os hexágonos ligados às elipses representam as máscaras tridimensionais associadas à respectiva classe. Quanto mais o ator A repetir uma dada unidade de ação ou emoção, mais exemplares de máscaras vão sendo associadas a essa classe.

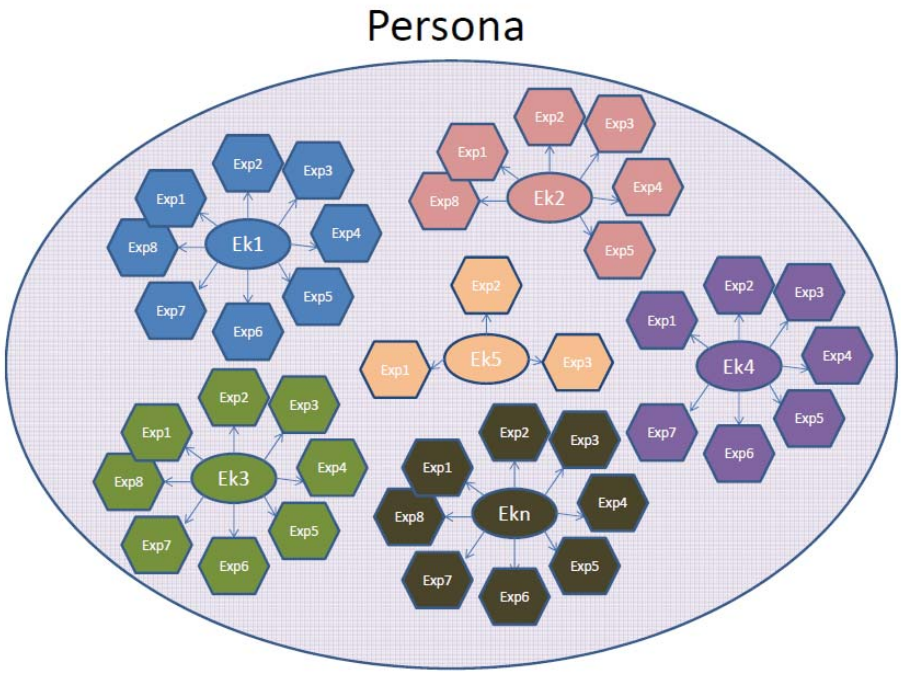


Figura 4.22: Estrutura de dados Persona.

Em um passo de pós-processamento, eliminam-se máscaras da mesma unidade de ação ou emoção que sejam muito semelhantes. Utiliza-se a distância de Hausdorff modificada [27] explicada na Seção 3.7 como medida de similaridade. Se essa distância entre duas máscaras da mesma unidade de ação ou expressão for menor que um limiar l_3 empiricamente obtido, uma delas é excluída da Persona. Para a boca, esse limiar corresponde a 5% da distância entre os cantos da boca. Para os olhos e sobrancelhas, 5% da distância entre os cantos dos olhos.

A manutenção de mais de uma máscara de controle associada à mesma unidade de ação ou emoção é justificada pela variabilidade de formas e intensidades com que o ator pode expressar a mesma unidade de ação ou emoção. Por exemplo, uma mesma pessoa pode sorrir de várias formas dependendo da circunstância. É importante manter na estrutura de dados as diversas formas de sorriso que o ator executou durante o treinamento. Resultados da construção da Persona podem ser vistos na Seção 5.1.

Uma vez criada a Persona do ator relacionada a um dado vídeo de entrada, utiliza-se agora essa estrutura de dados para animar o avatar, que pode ou não ter a mesma aparência do ator A. A próxima seção explica como se dá esse processo.

4.3 Utilização da Persona

Essa seção demonstra como a estrutura de dados Persona do ator A pode ser utilizada para animar o avatar de A a partir das informações obtidas pelo software de rastreamento dos pontos característicos da face do usuário U. O esquema da Figura 4.23 mostra resumidamente o processo de utilização da estrutura de dados Persona. Conforme essa figura, os pontos característicos da face do usuário devem ser rastreados no vídeo de entrada. De posse desses dados, procede-se para a classificação das unidades de ação da boca, olhos direito e esquerdo e sobrancelhas direita e esquerda. Sabendo-se a unidade de ação, escolhe-se na Persona a máscara de controle treinada com os movimentos faciais do ator adequada para cada um dos componentes faciais. Por fim, a combinação dessas máscaras passa a constituir parâmetros de animação para o avatar. A Figura 4.24 mostra esquematicamente esse processo e cada passo desse esquema será explicado no restante dessa seção.

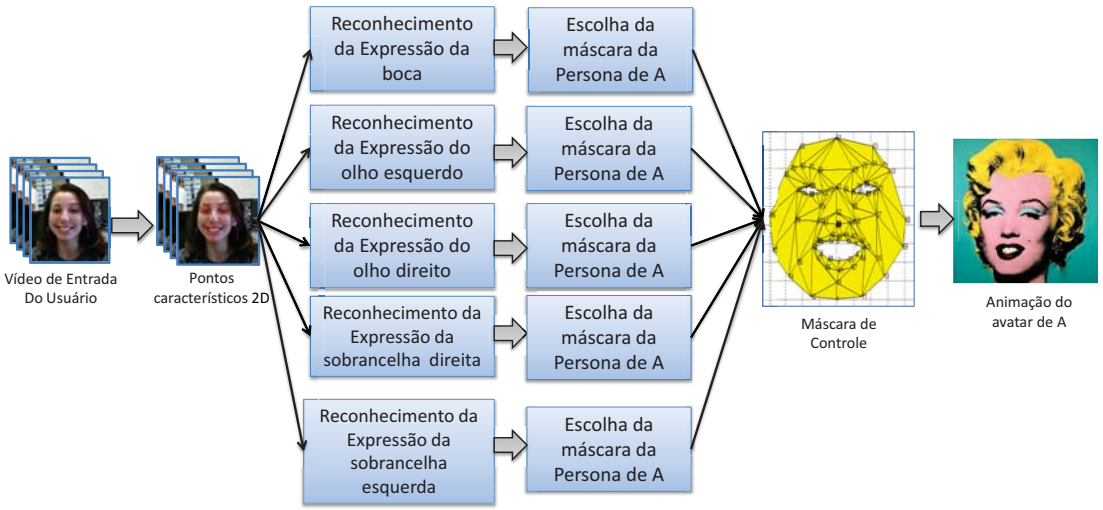


Figura 4.23: Processo de utilização da Persona.

De acordo com a Figura 4.24, o processo se inicia com o conjunto de dados Λ_I rastreado pelo *Live Driver* em um vídeo do usuário U. Primeiramente, a máscara de controle padrão $\mathcal{K}_0$ é rotacionada de acordo com o ângulo (β) fornecido pelo *Live Driver*. Esse ângulo refere-se à rotação da face em torno do eixo y e não pode ser determinado pela análise de Procrustes do conjunto Λ_I (em coordenadas de imagem) para registro nas coordenadas $(x_{\mathcal{K}_0}, y_{\mathcal{K}_0})$ da máscara de controle padrão. O objetivo dessa rotação é minimizar os efeitos da distorção da imagem devido à perspectiva durante o registro.

Após, os dados do conjunto Λ_I são registrados de acordo com a máscara padrão rotacionada, por meio da análise de Procrustes. Com a análise de Procrustes são obtidas a matriz de rotação $\mathbf{R}$ (2D), o vetor de translação $\mathbf{t}$ e o fator de escala s, calculados considerando-se seis pontos da face: os cantos externos dos olhos, a ponta do nariz, as narinas e o ponto intermediário do lábio superior. Com isso obtém-se o conjunto Λ' de pontos na mesma escala e posição relativa da máscara de controle rotacionada, através da Equação 4.9, aplicada a todos os dados do conjunto Λ_I:

$$(x_{\Lambda'_k}, y_{\Lambda'_k}) = s.(x_{\Lambda_{I_k}}, y_{\Lambda_{I_k}}).\mathbf{R} + \mathbf{t}. \tag{4.9}$$

Num último procedimento desse passo de registro, o conjunto Λ' é desrotacionado, aplicando-se o ângulo $(-\beta)$ e obtendo-se o conjunto de 64 pontos Λ, para que os pontos fiquem alinhados com a máscara padrão.

No início da utilização do sistema pelo usuário, é solicitado que o usuário U mantenha uma expressão facial neutra com mínimo ângulo de rotação em relação ao plano de aquisição da câmera. Nesse caso, armazena-se m conjuntos Λ e determina-se que o conjunto de pontos característicos da expressão neutra Λ_N será composto pela mediana de cada ponto dos m conjuntos Λ. O conjunto Λ é, então, dividido de acordo com os componentes da face de interesse, ou seja, boca (Λ_B), olho direito (Λ_{Od}), olho esquerdo (Λ_{Oe}), sobrancelha direita (Λ_{Sd}) e sobrancelha esquerda (Λ_{Se}). Cada conjunto é, após, deslocado para seu respectivo ponto âncora.

A segunda etapa descrita na Figura 4.24 é referente à extração das informações de entrada da Rede Neural Artificial. Para a obtenção desses dados, deve-se seguir um processo semelhante ao ilustrado na Figura 4.19 para cada componente da face separadamente: boca, olho direito, olho esquerdo, sobrancelha direita e sobrancelha esquerda. Resumidamente, dado que o componente facial da expressão analisada (boca, por exemplo) está na mesma escala e posição do componente facial da expressão neutra, calcula-se os deslocamentos horizontal e vertical, conforme as Equações 4.3. Os elementos dos vetores de deslocamento são então padronizados, de acordo com o conjunto de Equações 4.4. É feita então a mudança de base da Equação 4.7, mapeando-se a expressão para o subespaço principal. Esse novo vetor servirá de entrada para a Rede Neural Artificial (passo 3 da Figura 4.24).

Segue-se então a classificação das unidades de ação ou emoção presentes em cada subconjunto conjunto Λ_i, com $i = 1, ..., 5$. Caso a expressão de algum componente facial não seja identificada pelas RNAs de acordo com os critérios descrito na Seção 4.2.2 e ilustrados na Figura 4.20, esse quadro do vídeo não proverá informações para a animação do avatar. Nesse caso, o sistema de animação deverá interpolar as posições do último quadro cuja expressão foi reconhecida com sucesso até as posições dos vértices do próximo quadro cuja expressão seja reconhecida com sucesso. Nesse caso, a animação é feita por quadros chave.

Voltando à Figura 4.24, o procedimento de número 4 consiste em verificar na estrutura de dados Persona do ator A se a unidade de ação ou emoção identificada pelos classificadores foi incorporada por ocasião da sua construção. Se essa unidade de ação ou emoção existir na Persona, escolhe-se a

máscara que pertence a essa classe que melhor se adequa à ação do usuário. Para que essa seleção seja feita, é calculada a distância de Hausdorff modificada (descrita na Seção 3.7) entre os pontos pertencentes ao componente facial em questão da máscara e os pontos correspondentes do conjunto Λ. A máscara escolhida será aquela com menor DHM em relação aos pontos do usuário.

Caso a expressão não seja encontrada no passo 4 da Figura 4.24, verifica-se ainda se a classe obtida na etapa de classificação é composta (etapa 5). Relembrando, as classes compostas são designadas por c_1/c_2. Em caso afirmativo (a expressão do usuário é composta), verifica-se se há na Persona classe simples igual a expressão principal c_1 que o usuário está realizando. Caso essa classe não conste na Persona do ator A, o quadro de vídeo não retornará informações para a animação do avatar. Os parâmetros de animação nesse caso devem ser obtidos por interpolação pelo sistema de animação, conforme explicado anteriormente. Se a classe principal c_1 da classe composta do usuário (c_1/c_2) existir na Persona, procede-se à escolha da máscara de controle mais adequada à ação do usuário, utilizando-se novamente a DHM.

O fluxo da Figura 4.24 é repetido em cada quadro para cada componente facial i. Os parâmetros de animação assim obtidos são então, finalmente, informados ao sistema de animação facial como uma sequência de parâmetros para quadros-chave. Nessa sequência, indica-se o número do quadro q e os respectivos parâmetros. É importante ressaltar que pode haver quadros ausentes na sequência e, portanto, o sistema de animação deve prover o movimento contínuo entre os quadros-chave. Conforme mencionado anteriormente, isso pode ser obtido por meio de interpolação linear entre os parâmetros dos quadros informados na sequência de quadros-chave.

Resultados obtidos pelo modelo descrito serão apresentados no próximo capítulo.

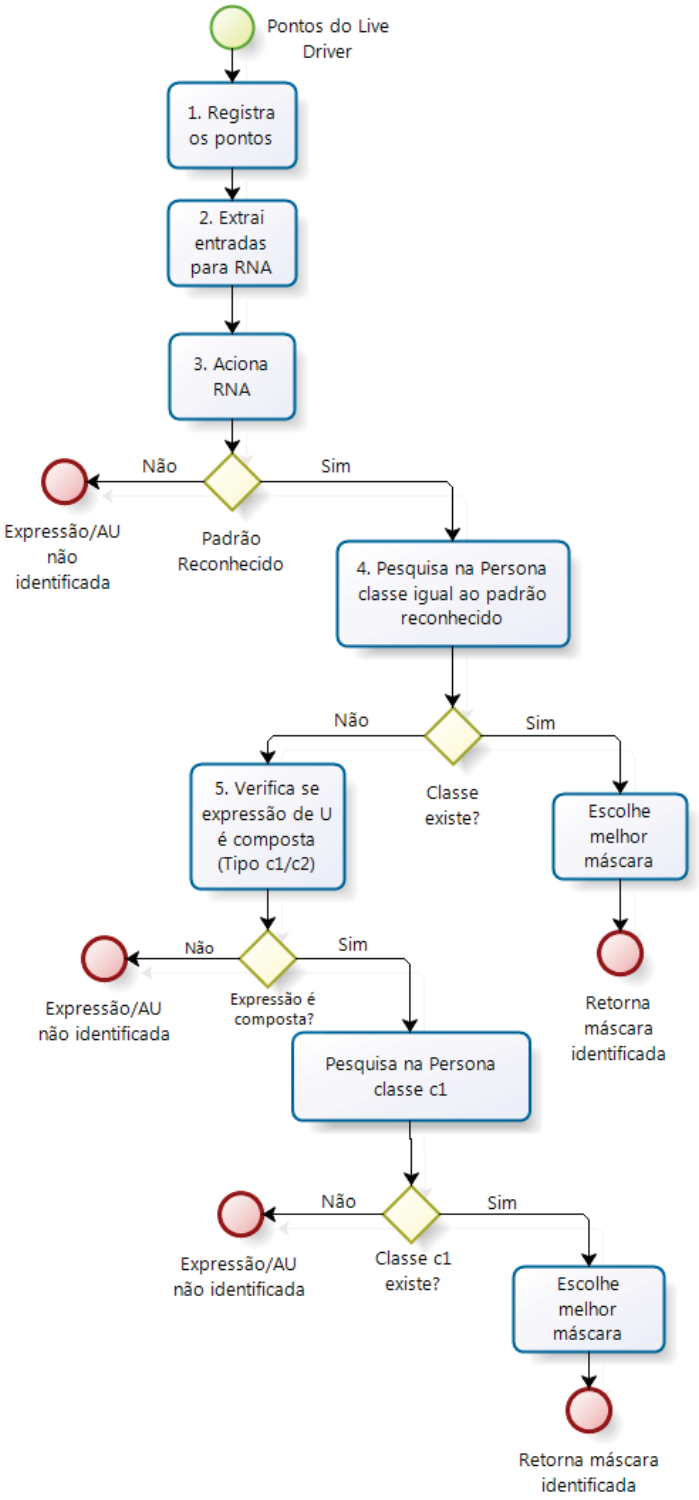


Figura 4.24: Processo de utilização da Persona.

5. RESULTADOS

Esse capítulo apresenta resultados obtidos com o modelo proposto nesta tese. Foram escolhidos dois casos de estudo para construção da Persona, a fim de exemplificar a aplicação do método descrito no capítulo anterior. Nesses casos de estudo, foram selecionados trechos de vídeo em que dois atores aparecem durante um razoável intervalo de tempo (maior que 2 min) tendo suas expressões faciais gravadas com boa proximidade da câmera e atuando de forma expressiva. O primeiro caso de estudo consiste na construção e utilização da Persona da cantora Sinéad O'Connor no videoclipe da música "Nothing Compares 2 U"[1]. No segundo caso de estudo, decidiu-se utilizar trechos do filme Questão de Honra (*A Few Good Men*)[2] nos quais o ator Jack Nicholson aparece atuando com expressões faciais marcantes e em *close*. As duas próximas seções mostram resultados obtidos com a construção e utilização das Personas nesses casos de estudo. A Seção 5.3 mostra a comparação visual entre os parâmetros gerados com a utilização das duas Personas para um vídeo espontâneo de uma usuária. Ao final desse capítulo serão discutidas questões de desempenho computacional e limitações da metodologia proposta.

5.1 Primeiro Caso de Estudo: Cantora Sinéad O'Connor

No primeiro caso de estudo, a Cantora Sinéad O'Connor interpreta a música "Nothing Compares 2 U" em um videoclipe, cujas imagens consistem principalmente no rosto da cantora em um fundo preto. Na maior parte desse videoclipe, a cantora interpreta a canção de forma emotiva, triste e às vezes expressando raiva. Foram removidos quadros que não mostravam o rosto da cantora, sendo, então, utilizados 6277 quadros. Esses quadros foram processados pelo *Live Driver*. Foi automaticamente reportado pela ferramenta o sucesso no rastreamento dos componentes faciais em 4056 quadros. A partir dos dados de rastreamento desses 4056 quadros, foi construída a estrutura Persona da cantora, de acordo com o procedimento descrito na Seção 4.2. Na análise apresentada nessa seção, será enfatizada a boca, por ela consistir em um componente facial mais crítico para os propósitos desse trabalho. Além de apresentar um maior número de unidades de ação ou emoção associadas, os movimentos são mais complexos que os dos olhos e sobrancelhas.

Primeiramente, serão apresentados resultados do processo de construção da Persona da cantora. A Figura 5.1 mostra classes de unidades de ação do FACS, representadas através de conjuntos de imagens, cujos resultados do rastreamento foram analisados, transformados em vetores de coeficientes dos componentes principais e classificados pela RNA do componente facial boca, treinada na fase de Pré-processamento. No canto inferior direito de cada conjunto de imagens é mostrada a imagem (ou as imagens) retirada do Bosphorus da(s) unidade(s) de ação ou emoção correspondente(s) a cada classe. São mostradas apenas oito das 34 classes referentes à boca da Persona da cantora,

[1] O videoclipe pode ser visto em `https://www.youtube.com/watch?v=iUiTQvT0W_0` (acessado em 10/07/2014).

[2] `http://www.imdb.com/title/tt0104257/` (acessado em 10/07/2014).

para fins de avaliação. Algumas dessas classes são simples (apenas uma unidade de ação ou emoção detectada pela RNA) e outras são compostas (do tipo c_1/c_2), conforme indicado textualmente nas imagens das classes.

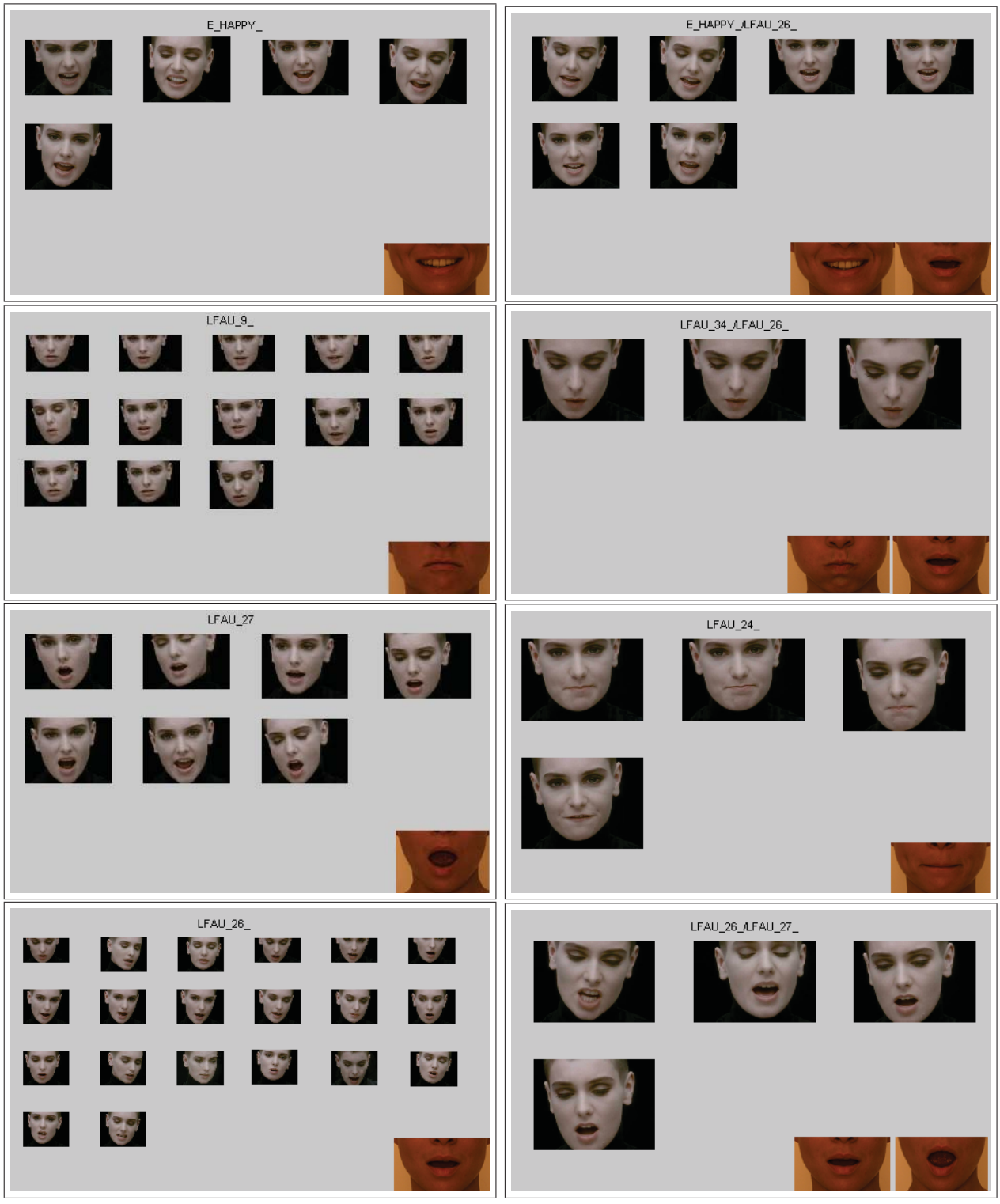


Figura 5.1: Imagens associadas a algumas das 34 classes geradas para a Persona da cantora Sinéad O'Connor. No canto inferior direito é mostrada a imagem (ou as imagens) retirada do Bosphorus da(s) unidade(s) de ação ou emoção correspondente(s) a cada classe

.

Algumas observações sobre os conjuntos de imagens de classes da Figura 5.1 se fazem necessárias. Pode-se perceber que o primeiro conjunto de imagens, classificadas como E_HAPPY, não expressam propriamente alegria na face da cantora. Porém, é preciso lembrar que, para a classificação das expressões, foram considerados apenas os deslocamentos dos lábios em relação à face neutra. Além disso, durante o treinamento, o único tipo de classe com abertura tanto lateral quanto vertical é a classe correspondente à emoção alegria (E_HAPPY), conforme imagens da Figura 3.22. Por esses dois motivos, essa classificação pode ser considerada correta, mesmo que a expressão da cantora não seja propriamente alegria. Uma extensão do banco de dados Bosphorus com mais combinações entre unidades de ação para treinamento das RNAs poderia resolver essa suposta inconsistência.

Considere agora os conjuntos de imagens da Figura 5.1 referentes às classes $LFAU\_27$ (primeira classe da terceira linha), $LFAU\_26$ (primeira classe da quarta linha) e $LFAU\_26/LFAU\_27$ (última classe apresentada). As duas primeiras são classes simples, indicando respectivamente, grande e média abertura vertical. A terceira classe é composta. No caso dessa última classe, ambas as unidades de ação (26 e 27) podem caracterizar a ação da cantora nas imagens pois representam situações intermediárias entre as $LFAUs$ 26 e 27. Como nos dados de treinamento não são claros os limites entre essas duas classes[3], ambas são aceitáveis.

A Figura 5.2 mostra, juntamente com as imagens que deram origem à classificação, as máscaras construídas em algumas classes da persona da cantora. Ao se analisar essa figura, pode-se observar que as imagens e máscaras da classe simples $(LFAU\_27)$ (caso em que o movimento de abertura da boca é somente para baixo) diferem-se da classe classificada como composta $(LFAU\_27/E\_HAPPY)$. Nessa última classe, pode-se notar que, conjuntamente como movimento de abertura da boca para baixo, há abertura lateral. Por isso, ambas as classificações são aceitáveis e é importante manter as duas na estrutura da Persona.

A Figura 5.3 exemplifica a escolha de máscaras da Persona da cantora Sinéad O'Connor, na etapa de utilização da Persona descrita na Seção 4.3. As imagens da coluna da esquerda mostram uma usuária cantando um trecho da música "Nothing Compares 2 U". Os visemas [4] selecionados para apresentação correspondem às letras sublinhadas da frase "'cos nothing compares to you", durante o refrão. Esses visemas foram classificados pela RNA da boca como unidades de ação $LFAU\_26$, $LFAU\_27$, $LFAU\_24$ e $LFAU\_9$, respectivamente. Essas unidades de ação são apresentadas nesta ordem, de cima para baixo, na Figura 5.3. Na terceira coluna dessa figura, são mostradas imagens correspondentes à máscara da Persona pertencente à mesma classe de unidades de ação ou emoção da usuária, cuja distância de Hausdorff modificada é mínima. Por exemplo, a unidade de ação executada pela usuária ao pronunciar as letras mp foi classificada como $LFAU\_24$. De acordo com os procedimentos descritos na Seção 4.3, verificou-se qual das máscaras pertencentes a essa classe na Persona tem menor DHM em relação aos pontos da boca da usuária rastreados pelo *Live*

[3] Os limites dessas classes não são bem definidos pois alguns indivíduos do Bosphorus são anatomicamente capazes de abrir a boca mais que outros, de forma que a abertura máxima de um pode ocasionar o mesmo deslocamento que a abertura média de outro.

[4] adaptação de *visemes* do inglês *Visual Phonemes*

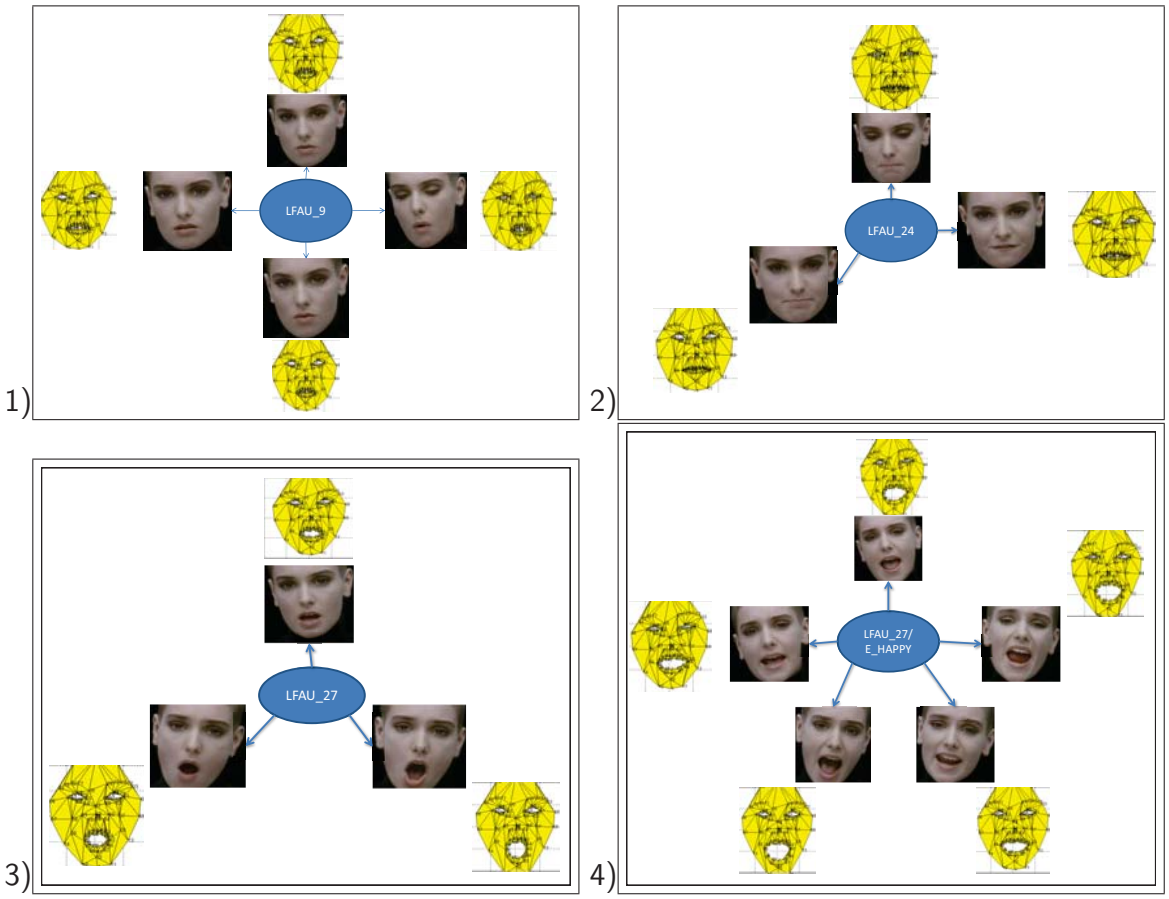


Figura 5.2: Quatro classes da Persona da cantora Sinéad O'Connor com imagens associadas e máscaras geradas. As unidades de ação $LFAU\_9$, $LFAU\_24$ e $LFAU\_27$ são caracterizadas como "Simples". Na quarta classe, as imagens contém expressões da boca que foram classificadas pelo método como compostas "$LFAU\_27/E\_HAPPY$".

Driver e registrados na máscara padrão. Apresentou-se então essa máscara na coluna da direita da Figura 5.3. Para fins de visualização, a máscara foi texturizada utilizando-se a imagem mostrada na coluna central. Essa texturização foi feita apenas para ilustração, sendo que regiões da testa estão escuras em virtude do fato de que os pontos do contorno superior da face (em coordenadas (u, v)) ficaram fora da imagem.

Com o objetivo de mostrar uma comparação quantitativa entre os parâmetros de animação obtidos *sem* a utilização da Persona em relação aos parâmetros de animação obtidos *com* a utilização da Persona, foram gerados três conjuntos de máscaras, considerando o mesmo vídeo da usuária cantando um trecho da música "Nothing Compares to You" (com R quadros) e o vídeo original da cantora (com Q quadros). Foram geradas:

- R máscaras $\mathcal{K}_U$ pelo processo descrito na Seção 4.2.3 para o vídeo da usuária U, sem utilização da Persona (essas máscaras foram geradas a partir dos pontos rastreados no rosto da usuária);

- R máscaras $\mathcal{K}_\mathcal{P}$ equivalentes ao mesmo trecho de vídeo, obtidas pela análise do movimento da usuária utilizando a Persona $\mathcal{P}$ da cantora Sinéad O'Connor (designada como atriz A);

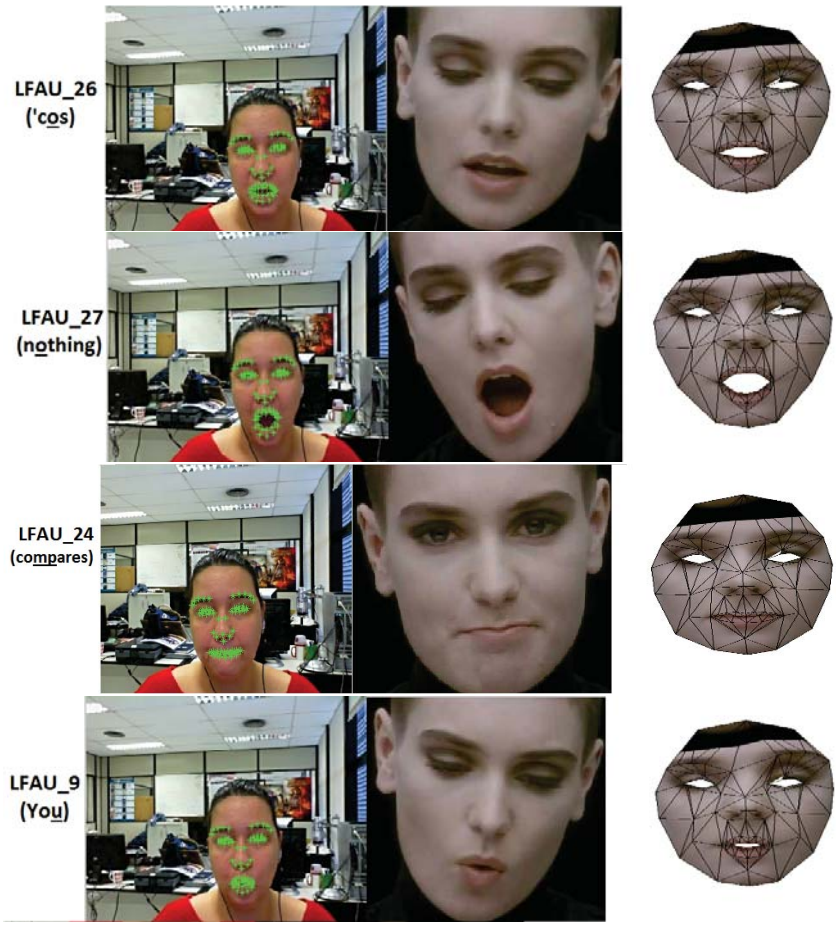


Figura 5.3: Unidades de ação da boca da usuária correspondentes na persona da cantora Sinéad O´Connor. De cima para baixo, são apresentadas as unidades de ação $LFAU\_26$, $LFAU\_27$, $LFAU\_24$ e $LFAU\_9$. Essas ações correspondem às letras sublinhadas nos textos à esquerda.

- Q máscaras $\mathcal{K}_A$ em cada quadro do vídeo original da atriz A e utilizadas na construção da Persona.

Com a comparação entre esses conjuntos de máscaras, pode-se avaliar se os parâmetros de animação gerados através da utilização da Persona ($\mathcal{K}_\mathcal{P}$) são mais semelhantes aos parâmetros referentes à ação da atriz ($\mathcal{K}_A$) do que os gerados sem utilização da Persona ($\mathcal{K}_U$), referentes somente à ação da usuária). Para essa comparação, foi utilizada a DHM entre os vértices das máscaras $\mathcal{K}_U$ e $\mathcal{K}_\mathcal{P}$ em relação às máscaras $\mathcal{K}_A$ em quadros correspondentes aos mesmos visemas (designadas por r no vídeo da usuária e por q no vídeo da atriz). A Figura 5.4 mostra imagens que deram origem às máscaras $\mathcal{K}_{A_q}$, $\mathcal{K}_{U_r}$ e $\mathcal{K}_{\mathcal{P}_r}$, correspondentes à letra "o" da palavra "nothing" no refrão. A imagem da esquerda dessa figura corresponde ao quadro q em que efetivamente a cantora vocalizou a letra "o" da palavra "nothing". A imagem central é referente à usuária pronunciando a mesma letra, no quadro r. Já a imagem da direita foi no processo de utilização da Persona da cantora para a expressão feita pela usuária.

Conforme mencionado no parágrafo anterior, foi calculada a distância de Hausdorff modificada H entre os vértices das máscaras $\mathcal{K}_{U_r}$ e os vértices da máscara $\mathcal{K}_{A_q}$, com r e q correspondentes

Figura 5.4: Imagens que deram origem às máscaras $\mathcal{K}_{U_r}$, $\mathcal{K}_{\mathcal{P}_r}$ e $\mathcal{K}_{A_q}$

.

a quadros da mesma sílaba da música. Da mesma forma, calculou-se a distância de Hausdorff modificada H entre as máscaras $\mathcal{K}_{\mathcal{P}_r}$ e $\mathcal{K}_{A_q}$. A Tabela 5.1 mostra os resultados dessas distâncias para alguns visemas. Para que se tenha uma noção de escala, foi medida a DHM entre as máscaras geradas em dois quadros consecutivos no vídeo da cantora, o que supostamente indica pouca diferenciação entre máscaras. O valor médio dessas distâncias, obtidas em 10 amostras aleatórias, foi de $H = 0,7$. Assim sendo, $H = 0,7$, a priori, representa um valor de DHM que indica pouca diferenciação.

Tabela 5.1: Comparação entre máscaras geradas a partir do vídeo do usuário não utilizando a Persona ($\mathcal{K}_{U_r}$) e utilizando a Persona $\mathcal{K}_{A_q}$.

Visema (letra considerada)	$H(\mathcal{K}_{U_r},\mathcal{K}_{A_q})$	$H(\mathcal{K}_{\mathcal{P}_r},\mathcal{K}_{A_q})$
‘Cos	18,0	15,4
Nothing	17,9	9,1
Compares	14,7	8,4
Compares	18,7	10,8
You	18,2	10,5

Analisando-se a Tabela 5.1, pode-se observar que o fato de utilizar as máscaras da Persona da cantora Sinéad O'Connor ($\mathcal{K}_{\mathcal{P}_r}$) diminui a DHM das máscaras da cantora geradas diretamente dos pontos de rastreamento do vídeo ($\mathcal{K}_{A_q}$), quando comparadas com as máscaras geradas diretamente pelo rastreamento dos pontos da face da usuária ($\mathcal{K}_{U_r}$). É importante salientar que as máscaras $\mathcal{K}_{A_q}$ e $\mathcal{K}_{\mathcal{P}_r}$ não são idênticas pois foram geradas com base em pontos rastreados em quadros diferentes de vídeo. Pela análise numérica, é possível concluir que $\mathcal{K}_{\mathcal{P}_r}$ é mais semelhante a $\mathcal{K}_{A_q}$ do que $\mathcal{K}_{U_r}$. Extrapolando essa consideração, pode-se afirmar então que os parâmetros de animação obtidos por meio do uso da Persona se assemelham mais à ação da cantora, que os parâmetros de animação obtidos diretamente pela análise do vídeo da usuária. Ao se observar a Figura 5.4, nota-se que a atuação da cantora no refrão é mais expressiva do que a atuação da usuária. Por isso, a imagem correspondente à máscara da Persona escolhida $\mathcal{K}_{\mathcal{P}_r}$ apresenta unidade de ação com menor intensidade do que a máscara correspondente à atuação da cantora no refrão $\mathcal{K}_{\mathcal{A}_q}$. Esse fato explica porque as distâncias de Hausdorff modificadas da Tabela 5.1 tem valores diferentes de zero.

Finalmente, para encerrar a análise de resultados do primeiro caso de estudo, serão apresentados alguns resultados de animação do avatar a partir das máscaras de controle da cantora Sinéad O'Connor guiadas pela performance da usuária U, mostrada na Figura 5.5. O objetivo da apresentação desses resultados é mostrar que as máscaras de controle obtidas por meio dos procedimentos descritos nesse trabalho podem guiar a animação de faces de avatares. Entretanto, até a data da escrita dessa tese, o sistema de animação, que é responsável pela deformação de vértices da malha da face do avatar não coincidentes com os vértices da máscara de controle de animação, não se encontra completamente finalizado. Por esse motivo não foram realizadas avaliações qualitativas com pessoas observando o resultado das animações. Mesmo no estágio atual do sistema de animação, entretanto, é possível verificar a correspondência entre as expressões da usuária, da cantora e do avatar.

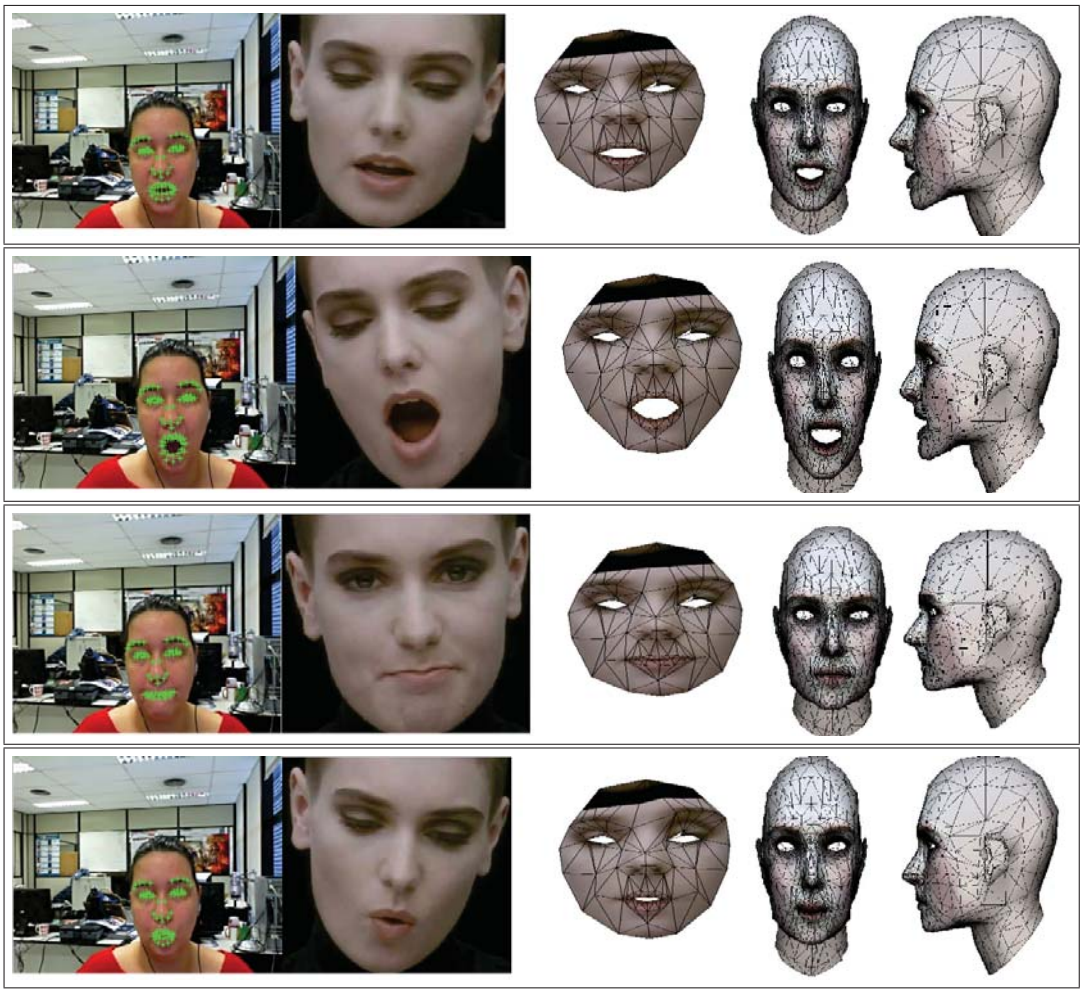


Figura 5.5: Resultado da animação do avatar (imagens à direita) por meio da máscara de controle obtida pela utilização da Persona. A usuária da esquerda está executando ações classificadas como $LFAU\_26$, $LFAU\_27$, $LFAU\_24$ e $LFAU\_9$, de cima para baixo.

A próxima seção descreve avaliações semelhantes considerando o segundo caso de estudo.

5.2 Segundo Caso de Estudo: Ator Jack Nicholson

O segundo caso de estudo desse trabalho consiste na construção e utilização da persona do ator Jack Nicholson no filme "Questão de Honra"[5]. Nesse caso, a performance do ator em duas cenas foi analisada. Na primeira cena, é focado o rosto do ator durante um diálogo. Na segunda cena, o personagem está dando testemunho em um julgamento. A emoção principal mostrada pelo ator nessa última cena é ira. Foram utilizados ao todo 2239 quadros, tendo sido reportado sucesso no rastreamento do *Live Driver* em 2013 quadros. A Figura 5.6 mostra imagens associadas a algumas unidades de ação ou emoção do ator Jack Nicholson no processo de construção da Persona.

Assim como realizado no primeiro caso de estudo, solicitou-se que um usuário repetisse a atuação do ator Jack Nicholson em um trecho do vídeo usado para construir a Persona. Foi solicitado que o usuário repetisse um trecho do testemunho do personagem no filme, o qual se inicia com a frase "*You can't handle the truth*", dita de forma enfática e expressando raiva. Imagens de alguns fonemas visuais ou visemas dessa frase podem ser vistas na Figura 5.7. Na tabela mostrada nessa figura, cada linha representa um visema. Na coluna "Imagem do Usuário", aparece o usuário U pronunciando as letras sublinhadas na primeira coluna. Na coluna "Imagem Correspondente da Persona", aparece a imagem da Persona do ator Jack Nicholson escolhida de acordo com a Seção 4.3. Na coluna "Expressão feita pelo ator" é mostrada a imagem efetivamente realizada pelo ator ao pronunciar os fonemas sublinhados na referida frase no filme "Questão de Honra". As máscaras mostradas serão explicadas a seguir.

Assim como no primeiro caso de estudo, foram construídas máscaras $\mathcal{K}_{U_r}$ adaptadas ao rosto do usuário, sem utilização da Persona, para compará-las com as máscaras $\mathcal{K}_{\mathcal{P}_r}$ escolhidas na Persona referentes ao quadro r do vídeo do usuário U. Para essa comparação, foi calculada a DHM entre a máscara $\mathcal{K}_{\mathcal{A}_q}$ associadas ao quadro q do vídeo do ator utilizado na construção da persona e as máscaras $\mathcal{K}_{U_r}$ e $\mathcal{K}_{\mathcal{P}_r}$. Assim, na Figura 5.7, as máscaras $\mathcal{K}_{\mathcal{P}_r}$ estão associadas às imagens da coluna "Imagem Correspondente da Persona" e as máscaras $\mathcal{K}_{\mathcal{A}_q}$ estão associadas às imagens da coluna "Expressão feita pelo ator". Os valores de algumas dessas distâncias estão na Tabela 5.2. Foram utilizados nessa tabela os visemas referentes à letra "Y" na palavra "You", "A" na palavra "Can't" e "A" na palavra "Walls".

Tabela 5.2: Comparação entre máscaras geradas a partir do vídeo do usuário não utilizando a Persona ($\mathcal{K}_{U_r}$) e utilizando a Persona $\mathcal{K}_{A_q}$.

Visema (letra considerada)	$H(\mathcal{K}_{U_r},\mathcal{K}_{A_q})$	$H(\mathcal{K}_{\mathcal{P}_r},\mathcal{K}_{A_q})$
Y̲ou	21,68	18,0
Ca̲n't	31,26	20,1
Wa̲lls	12,15	3,95

A análise da Tabela 5.2 mostra que as distâncias de Hausdorff modificada $H(\mathcal{K}_{U_r},\mathcal{K}_{A_q})$ entre as máscaras $\mathcal{K}_{U_r}$ e as máscaras do ator $\mathcal{K}_{A_q}$ são maiores que as distâncias $H(\mathcal{K}_{\mathcal{P}_r},\mathcal{K}_{A_q})$ entre as

[5] *A Few Good Men* - Columbia Pictures Corporation - 1992

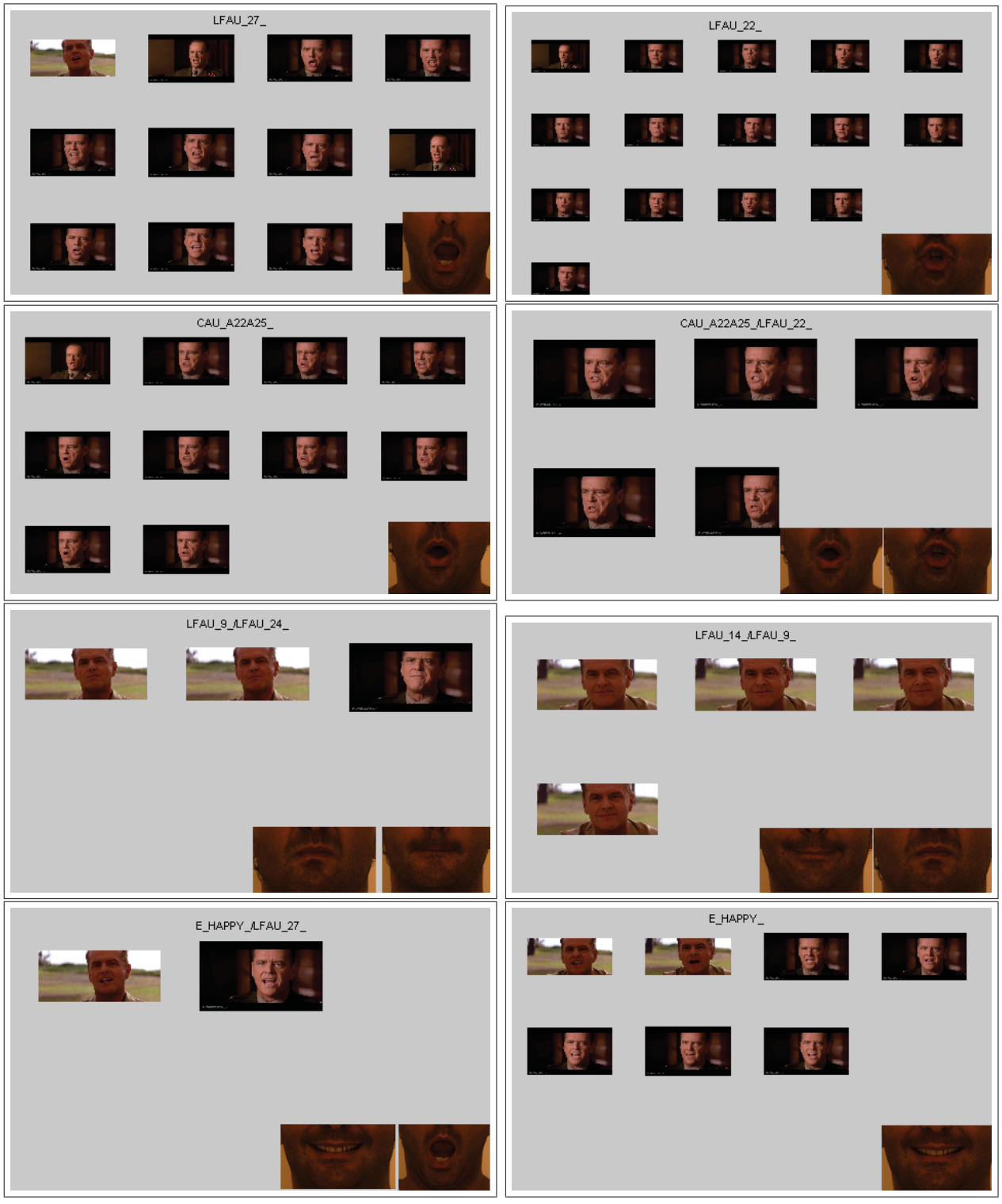


Figura 5.6: Imagens associadas a algumas classes da Persona do ator Jack Nicholson no filme "Questão de Honra".

máscaras da Persona escolhidas $\mathcal{K}_{\mathcal{P}_r}$ e as máscaras do ator $\mathcal{K}_{A_q}$. Assim, embora não tenham sido escolhidas máscaras idênticas à ação do ator, as medidas indicam que as máscaras escolhidas via utilização da Persona são mais semelhantes à ação do ator do que sem utilização da Persona, para os visemas considerados. Há que se considerar que o usuário não é um ator profissional e sua forma

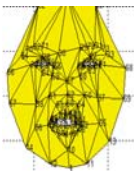
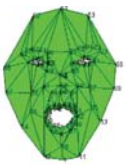
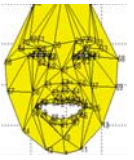
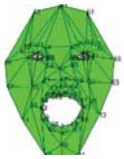
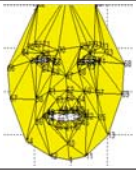
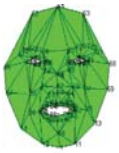
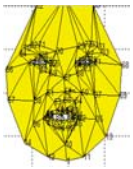
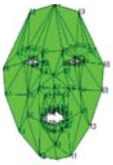
Visema	Imagem do Usuário	Imagem Correspondente da Persona	Expressão feita pelo Ator	$\mathcal{K}_{U_r}$	$\mathcal{K}_{\mathcal{P}_r}$	$\mathcal{K}_{A_q}$
You						
Can't						
Handle						
Truth						

Figura 5.7: Tabela para a comparação entre a ação do usuário, a expressão do ator escolhida na Persona e a expressão efetivamente realizada pelo ator Jack Nicholson. Nessa tabela, são mostradas também as máscaras $\mathcal{K}_{\mathcal{P}_r}$ obtidas pela utilização da Persona e as máscaras $\mathcal{K}_{\mathcal{A}_q}$ obtidas diretamente do vídeo do ator.

de se expressar não continha uma manifestação de raiva tão intensa quanto na atuação do ator Jack Nicholson.

5.3 Comparação das Personas dos Casos de Estudo 1 e 2

Para encerrar a seção de resultados, a Figura 5.8 mostra imagens obtidas com a utilização da Persona a partir de um vídeo espontâneo de uma usuária. O objetivo dessa figura é apresentar resultados obtidos quando o usuário não interpreta cenas semelhantes às utilizadas na construção da Persona. Assim sendo, nesse vídeo de entrada, a usuária aparece falando, sorrindo e não há roteiro pré-definido. Cada linha da Figura 5.8 mostra a classe detectada pelas RNAs no quadro do vídeo mostrado na coluna 1. A coluna 2 mostra a imagem escolhida nessa classe da Persona da contora Sinéad O'Connor. A coluna 3 mostra as máscaras associadas às imagens da coluna 2, pertencentes também à Persona da cantora Sinéad O'Connor. A coluna 4 mostra as imagens do ator Jack Nicholson associada às classes de cada linha pertencentes à Persona do ator. Já a coluna 5 mostra as máscaras escolhidas em cada classe na Persona do ator Jack Nicholson, correspondentes

às imagens da coluna 4. É importante salientar que, apesar da usuária não interpretar o mesmo texto que os atores, as máscaras selecionadas na Persona são visualmente adequadas às suas ações.

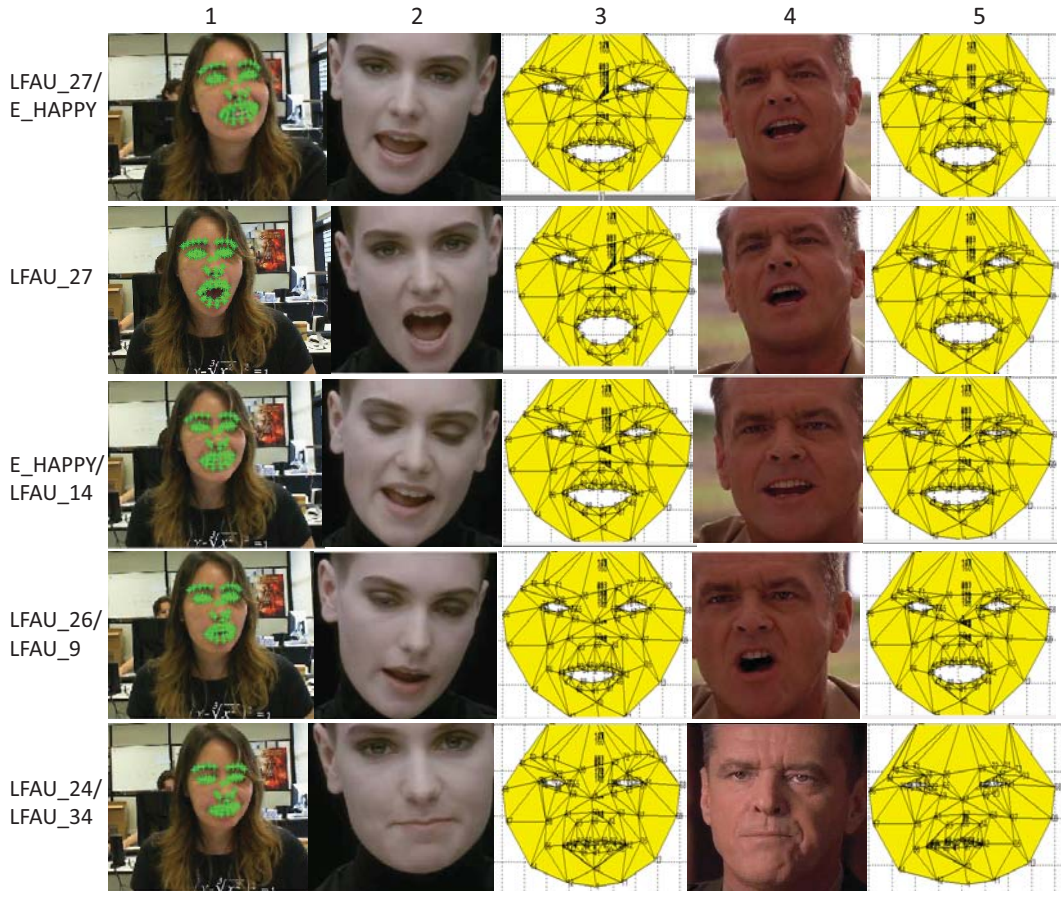


Figura 5.8: Resultados da utilização da Persona para um vídeo de entrada espontâneo de uma usuária. As colunas 2 e 3 mostram respectivamente a imagem e a máscara da Persona da cantora Sinéad O'Connor atribuídas pelo método à cada imagem da usuária na coluna 1. Igualmente, as colunas 4 e 5 mostram as imagens e as máscaras da Persona do ator Jack Nicholson escolhidas.

5.4 Desempenho Computacional do Método

Parte do desenvolvimento do método é feito em MatLab e parte é feito em C++. Visto que o objetivo é a prova do conceito de geração/utilização da Persona e não um aplicativo ou software em si, a análise do desempenho não visou tempo real. A aplicação principal no momento é a pós-produção de vídeo de usuários e atores, seja para a geração ou para a utilização da Persona. Assim, o método não funciona em tempo real. São necessários em torno de 2 segundos para a geração de um máscara por quadro, calculado em Notebook padrão, com 3GB de Memória. Se for gerado um protótipo completo em C++, medidas de desempenho mais significativas poderão ser realizadas no futuro.

Quanto à utilização da memória, a Persona da cantora Sinéad O'Connor, por exemplo, gerou 34 classes, tendo cada uma delas em média 10 máscaras. Assim, 340 máscaras com 86 vértices não

representa grande custo de utilização de memória.

5.5 Limitações da Metodologia Proposta

O problema apresentado nessa tese é inovador pelo que se pôde perceber na bibliografia pesquisada e os resultados mostrados nas Seções 5.1 e 5.2 mostram que a metodologia descrita dá indícios de que é possível resolvê-lo. Evidentemente, há limitações nesse protótipo inicial que são ocasionadas pontualmente em diversas etapas. As principais limitações podem ser enumeradas conforme segue:

1. Inferência sobre a terceira coordenada das máscaras de controle;
2. Erros de classificação devidos à rotação da face em torno do eixo x;
3. Dependência da escolha da expressão neutra.

Quanto ao item número 1, observou-se que o fato de serem escolhidas diferentes pessoas P do banco de dados Bosphorus para a inferência de deslocamentos no eixo z pode levar a instabilidades na animação. Esse problema poderia ser resolvido utilizando-se os *Morphable Models* propostos por Blanz e Vetter [5]. Essa técnica foi proposta para reconstrução de modelos geométricos tridimensionais de faces a partir de imagens. Tais modelos devem ser treinados a partir de bancos de dados tridimensionais. Não houve tempo hábil para aplicação dessa técnica com base nos dados do Bosphorus.

A limitação de número 2 é ocasionada pela deformação aparente dos componentes faciais devido ao fato da face não estar contida no plano da imagem. Com isso, o tamanho aparente dos lábios, por exemplo, muda em relação aos demais pontos da face, conforme pode ser observado na Figura 5.9. O processo de registro não resolve eficientemente esse problema e a mudança aparente é percebida como deslocamentos do contorno dos lábios em relação à face neutra. Investigações mais rigorosas devem ser feitas em trabalhos futuros para melhorar a classificação nesses casos. A Figura 5.9 mostra um caso de erro de classificação por esse motivo. Os pontos do contorno dos lábios foram classificados como $LFAU\_34$ (que poderia ser descrito como um biquinho), apesar da cantora estar com a boca neutra. Esse erro decorre por causa do estreitamento aparente dos lábios em relação à expressão neutra utilizada para o cálculo dos deslocamentos, mostrada na imagem da direita. A imagem central mostra um exemplo de indivíduo P do Bosphorus executando essa unidade de ação. Esse foi o indivíduo utilizado para construção da máscara 3D.

A terceira limitação apontada decorre da sensibilidade do método à máscara referente à expressão neutra, determinada de acordo com o processo descrito na Seção 4.3. Como o movimento dos lábios é variado e sutil e todo o processo depende do deslocamento dos pontos em relação à face neutra, a determinação da máscara referente à essa expressão é crucial. Se o usuário ficar com a boca entreaberta durante a aquisição dos quadros para construção da máscara neutra, por

Figura 5.9: Erro de classificação devido à Rotação da face. A unidade de ação dos lábios da imagem da esquerda foi classificada como $LFAU\_34$. A imagem central mostra um indivíduo do Bosphorus executando essa ação. Esse erro decorre, provavelmente, pela redução aparente da espessura dos lábios em relação à face neutra, mostrada na imagem da direita.

exemplo, quando ele fechá-la a unidade de ação será caracterizada como compressão dos lábios (provavelmente, $LFAU\_24$). Se, por outro lado, comprimir um pouco os lábios durante a construção da máscara neutra, quando relaxá-los a RNA classificará a expressão como abertura leve da boca (provavelmente, $LFAU\_25$).

Tais limitações, entretanto, podem ser resolvidas com maiores investigações no futuro. A próxima seção indica algumas ações planejadas para os próximos passos e apresenta considerações finais.

6. CONSIDERAÇÕES FINAIS

Esse trabalho apresentou uma metodologia para aprendizado e transferência do estilo de movimento facial (Persona) de atores para a animação de avatares. Por meio dessa metodologia pode-se guiar a animação das faces de avatares através da atuação de usuários diferentes do ator. Dessa forma, o avatar poderá expressar as unidades de ação ou emoções do usuário, porém com o estilo de movimento do ator.

O protótipo desenvolvido como prova de conceito utiliza como informações de entrada os dados de rastreamento de pontos de controle da face por ferramentas disponíveis comercialmente. Para classificação de unidades de ação ou emoção foram utilizadas redes neurais artificiais. A saída do protótipo se constitui da máscaras de controle para animação facial.

Na análise de resultados, foram apresentados casos de estudo para a construção da Persona de dois atores, seguida de sua utilização por usuários. Os resultados obtidos mostram eficiência no uso de classificadores automáticos de unidades de ação e emoções para construção e utilização da Persona, para movimentos da boca, olhos e sobrancelhas.

O problema de pesquisa apresentado mostrou-se uma área inovadora, de acordo com a revisão bibliográfica realizada. Por esse motivo, considerou-se que os resultados obtidos são válidos como prova de que é possível aprender e utilizar a Persona de Atores. Entretanto, uma série de investigações futuras deve ser realizada a fim de aprimorar cada etapa do processo, conforme a próxima Seção.

6.1 Trabalhos Futuros

Ao final desse trabalho pôde-se perceber uma série de investigações a serem realizadas para aprimorar os resultados obtidos ou, ainda, estender o método para incluir mais características à Persona. Essas investigações são listadas a seguir:

- Extensão do banco de dados: com a aquisição de um Scanner 3D pelo Centro de Pesquisas em Visualiação Avançada da PUCRS (CPVA), pode-se estender o banco de dados de faces tridimensionais. Isso permitirá a obtenção de nuvens de pontos mais densas com um maior número de exemplares e combinações de unidades de ação. A Figura 6.1 mostra uma face digitalizada por esse scanner;

- Investigação da utilização de máscaras de controle com maior número de pontos: com o desenvolvimento paralelo a esse trabalho do sistema de animação facial do laboratório de Simulação de Humanos Virtuais, pôde-se perceber a necessidade de um controle mais fino da animação. Por isso, um passo importante nos trabalhos futuros é aumentar o número de vértices da máscara de controle, atualmente com 86 vértices;

Figura 6.1: Imagens frontal e lateral de nuvem de pontos de uma face digitalizada pelo scanner 3D do CPVA.

- As rugas são detalhes importantes na expressão da face. Pode-se utilizar algoritmos de visão computacional que, combinados com o *Live Driver*, permitam detecção dessas regiões para informar a presença de rugas nas faces dos atores ao sistema de animação. As rugas mais fáceis de detectar e que representam importante papel na animação facial são aquelas que ligam os cantos da boca às laterais do nariz e as que ficam entre as sobrancelhas. Conforme apresentado na Seção 2.2, a detecção da presença dessas rugas deve inclusive auxiliar na classificação de unidades de ação ou emoção dos atores e usuários;

- Devem ser realizados testes de outros métodos de classificação mais recentes, utilizados em trabalhos citados na Seção 2.2, como combinações de classificadoreds fracos (*Adaboost* com *bootstrapping*);

- Aprendizado da dinâmica do movimento entre unidades de ação ou emoção: a análise das trajetórias das representações dos componentes faciais no subespaço principal pode levar à extensão da Persona para incluir a forma com que os atores mudam de uma dada unidade de ação ou emoção para outra. Assim, não só serão aprendidas as formas como cada ação é executada, mas também a transição de uma ação a outra. A Figura 6.2 mostra dois exemplos de transições entre ações e a respectiva representação de cada ação intermediária. No caso da cantora Sinéad O'Connor, a transição mostrada é de boca neutra para $LFAU\_27$. Repare que essa transição pode ser representada por uma trajetória do ponto vermelho no subespaço principal. Já o conjunto de imagens inferior mostra a transição da unidade de ação $LFAU\_34$ para a emoção E_HAPPY. A trajetória do ponto vermelho no subespaço principal representa a forma individual com que esse ator realiza essa transição específica. Acredita-se que esse tema contém muita inovação e deverá ser abordado no futuro.

- O método para construção de máscaras 3D, entretanto, mostrou-se um tanto primitivo. Ele pode, entretanto, ser aprimorado com o uso de *Morphable Models* [5]. Essa técnica é utilizada para reconstrução de modelos tridimensionais a partir de fotografias.

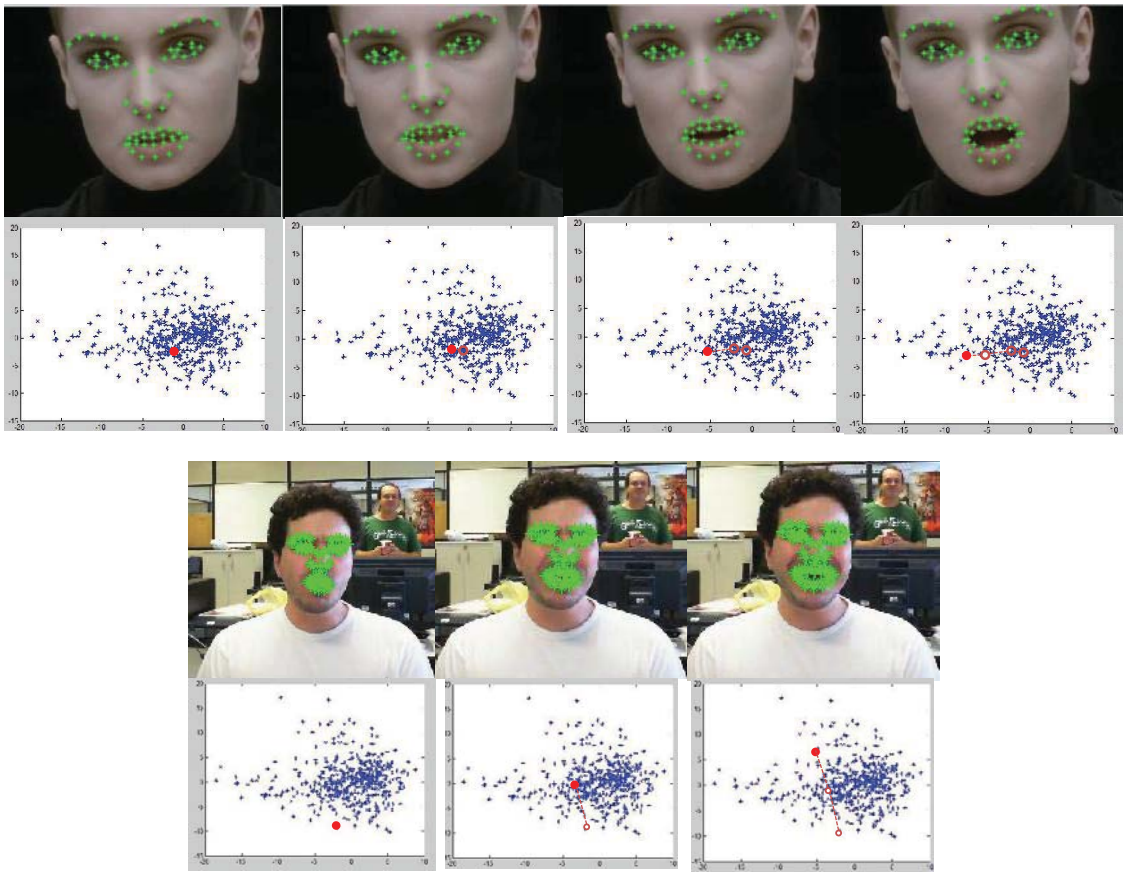


Figura 6.2: Duas sequência de imagens de transição de ação da boca. As imagens mostram as sequências de ações dos atores com os pontos rastreados. Os gráficos apresentam a representação dos pontos da boca nos dois primeiros componentes principais obtidos via PCA. Os pontos em azul são correspondentes às unidades de ação da boca dos indivíduos do Bosphorus, enquanto os pontos vermelhos correspondem à ação dos atores nas fotos.

- Realizar uma pesquisa qualitativa sobre a percepção, por parte das pessoas, da utilização da Persona em avatares.

Por fim, acredita-se que essa tese cumpriu seu papel em propor um método para aprender e utilizar o estilo de movimento de atores em avatares.

Bibliografia

[1] Ahlberg, J. Candide-3 – an updated parameterized face. Tech. Rep. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linkoping University, Sewden, 2001.

[2] Ahlberg, J. Candide 3: An updated parameterized face. Tech. rep., Dept. of Electrical Engineering, Linkoping University, Sweden, 2001.

[3] Ari, I., Uyar, A., and Akarun, L. Facial feature tracking and expression recognition for sign language. In *23rd Int. Symp on Computer and Information Science* (2008), IEEE Computer Society, pp. 1–6.

[4] Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] Blanz, V., and Vetter, T. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1999), SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., pp. 187–194.

[6] Brand, M., and Hertzmann, A. Style machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2000), SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., pp. 183–192.

[7] Buenaposada, J., noz, E. M., and Baumela, L. Efficient ilumination independent appearence-based face tracking. *Image and Vision Computing* (2008).

[8] Cao, C., Weng, Y., Lin, S., and Zhou, K. 3d shape regression for real-time facial animation. *ACM Trans. Graph. 32*, 4 (July 2013), 41:1–41:10.

[9] Chiu, C.-C., and Marsella, S. A style controller for generating virtual human behaviors. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3* (Richland, SC, 2011), AAMAS '11, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1023–1030.

[10] Chu, W.-S., de la Torre, F., and Cohn, J. Selective transfer machine for personalized facial action unit detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (June 2013), pp. 3515–3522.

[11] Chuang, C.-F., and Shih, F. Y. Recognizing facial action units using independent component analysis and support vector machine. *Pattern Recognition 39*, 9 (2006), 1795 – 1798.

[12] Cootes, T. F., Edwards, G. J., and Taylor, C. J. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23* (2001), 681–685.

[13] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. Active shape models - their training and application. *Computer Vision and Image Understanding 61* (1995), 38–59.

[14] Cristinacce, D., and Cootes, T. F. Facial feature detection and tracking with automatic template selection. In *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 429–434.

[15] Darwin, C. *The Expression of the Emotions in Man and Animals*. John Murray, London, 1872.

[16] DeCarlo, D., and Metaxas, D. Deformable model-based shape and motion analysis from images using motion residual error. In *Proceedings of the Sixth International Conference on Computer Vision* (Washington, DC, USA, 1998), ICCV '98, IEEE Computer Society, pp. 113–119.

[17] Dubuisson, M.-P., and Jain, A. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on* (Oct 1994), vol. 1, pp. 566–568 vol.1.

[18] Ekman, P., Friesen, W. V., and Hager, J. C. *The Facial Action Coding System*. Weidenfeld & Nicolson, 2002.

[19] Ekman, P., and Rosenberg, E. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using Facial Action Coding System (FACS)*, 2nd ed. Oxford University Press, New York, 2004.

[20] Etemad, S., and Arya, A. Recognition and re-synthesis of 3d human motion with personalized variations. In *Multimedia Computing and Systems, 2009. ICMCS '09. International Conference on* (April 2009), pp. 106–111.

[21] Fasel, B., and Luettin, J. Automatic facial expression analysis: a survey. *Pattern Recognition 36* (2002), 259–275.

[22] Gonzalez, I., Sahli, H., Enescu, V., and Verhelst, W. Context-independent facial action unit recognition using shape and gabor phase information. In *Affective Computing and Intelligent Interaction*, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., vol. 6974 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2011, pp. 548–557.

[23] Haj, M. A., Orozco, J., Gonzalez, J., and J. J. Villanueva, J. Automatic face and facial features initialization for robust and accurate tracking. In *19th International Conference on Pattern Recognition* (2008), IEEE Computer Society, pp. 1–4.

[24] Hamm, J., Kohler, C. G., Gur, R. C., and Verma, R. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods 200*, 2 (2011), 237 – 256.

[25] Haykin, S. *Redes Neurais: Princípios e Prática*, 2 ed. Bookman, 2001.

[26] Hirsch, M. *Differential Topology*. Graduate Texts in Mathematics. Springer, 1976.

[27] Hossain, M., Dewan, M., Ahn, K., and Chae, O. A linear time algorithm of computing hausdorff distance for content-based image analysis. *Circuits, Systems, and Signal Processing 31*, 1 (2012), 389–399.

[28] houh Chen, C., Härdle, W., and Unwin, A. *Handbook of Data Visualization*. American Psychological Association, 2008.

[29] Kanade, T., Cohn, J. F., and Tian, Y. Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (2000), pp. 46–53.

[30] Kendall, D. G. A survey of the statistical theory of shape. *Statistical Science* (1989), 87–99.

[31] Kim, J.-B., Hwang, Y., Bang, W.-C., Lee, H., Kim, J., and Kim, C. Real-time realistic 3d facial expression cloning for smart tv. In *Consumer Electronics (ICCE), 2013 IEEE International Conference on* (Jan 2013), pp. 240–241.

[32] Kotsia, I., Zafeiriou, S., and Pitas, I. Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition 41*, 3 (2008), 833 – 851. Part Special issue: Feature Generation and Machine Learning for Robust Multimodal Biometrics.

[33] Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), IEEE, pp. 3361–3368.

[34] Lee, C.-S., and Samaras, D. Analysis and synthesis of facial expressions using decomposable nonlinear generative models. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (March 2011), pp. 847–852.

[35] Lee, T. S. Image representation using 2d gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 18*, 10 (Oct 1996), 959–971.

[36] Li, H., Yu, J., Ye, Y., and Bregler, C. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics 32*, 4 (July 2013).

[37] Li, Y., Wang, S., Zhao, Y., and Ji, Q. Simultaneous facial feature tracking and facial expression recognition. *Image Processing, IEEE Transactions on 22*, 7 (July 2013), 2559–2573.

[38] Li, Z., Chen, J., Chong, A., Yu, Z., and Schraudolph, N. N. Using stochastic gradient-descent scheme n appearance model based face tracking. In *Proc. Intl. Workshop Multimedia Signal Processing (MMSP)* (Cairns, Australia, 2008), IEEE.

[39] Liu, G., Pan, Z., and Lin, Z. Style subspaces for character animation. *Comput. Animat. Virtual Worlds 19*, 3-4 (Sept. 2008), 199–209.

[40] Lowe, D. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on* (1999), vol. 2, pp. 1150–1157 vol.2.

[41] Lucas, B. D., and Kanade, T. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2* (San Francisco, CA, USA, 1981), Morgan Kaufmann Publishers Inc., pp. 674–679.

[42] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (Berkeley, Calif., 1967), University of California Press, pp. 281–297.

[43] Mahalanobis, P. C. On the generalised distance in statistics. In *Proceedings National Institute of Science, India* (Apr. 1936), vol. 2, pp. 49–55.

[44] Mahoor, M., Zhou, M., Veon, K. L., Mavadati, S., and Cohn, J. Facial action unit recognition with sparse representation. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (March 2011), pp. 336–342.

[45] Maya, A. Autodesk maya 2014. `http://download.autodesk.com/global/docs/maya2014/en_us/index.html?url=files/Blend_Shape_deformer.htm,topicNumber=d30e347338`, 2014.

[46] Mehrabian, A. Communication without words. *Psychology Today* (1968), 53–56.

[47] Min, J., Liu, H., and Chai, J. Synthesis and editing of personalized stylistic human motion. In *Proceedings of the 2010 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games* (New York, NY, USA, 2010), I3D '10, ACM, pp. 39–46.

[48] Minsky, M. L., and Papert, S. *Perceptrons: An Introduction to Computational Geometry*, expanded ed. ed. MIT Press, Cambridge Mass., 1988.

[49] Mpiperis, I., Malassiotis, S., and Strintzis, M. G. Bilinear elastically deformable models with application to 3d face and facial expression recognition. In *FG* (2008), pp. 1–8.

[50] Neff, M., and Kim, Y. Interactive editing of motion style using drives and correlations. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2009), SCA '09, ACM, pp. 103–112.

[51] Ong, E.-J., and Bowden, R. Robust facial feature tracking using shape-constrained multiresolution-selected linear predictors. *IEEE Trans. Pattern Anal. Mach. Intell. 33*, 9 (2011), 1844–1859.

[52] Pan, W., and Torresani, L. Unsupervised hierarchical modeling of locomotion styles. In *Proceedings of the 26th Annual International Conference on Machine Learning* (New York, NY, USA, 2009), ICML '09, ACM, pp. 785–792.

[53] Pandzic, R., and Fochheimer, R. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. Wiley, John and Sons, 2002.

[54] Pantic, M., and Rothkrantz, L. J. M. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 34*, 3 (June 2004), 1449–1461.

[55] Papageorgiou, C., Oren, M., and Poggio, T. A general framework for object detection. In *Computer Vision, 1998. Sixth International Conference on* (Jan 1998), pp. 555–562.

[56] Parent, R. *Computer animation: algorithms and techniques*. Elsevier, 2012.

[57] Pei, Y., and Zha, H. Transferring of speech movements from video to 3d face space. *IEEE Transactions on Visualization and Computer Graphics* (2007), 58–69.

[58] Pianigiani, O. *Vocabolario etimologico della lingua italiana*. Societa editrice Dante Alighieri di Albrighi, 2012.

[59] Queiroz, R. B. Geraï¿½ï¿½o de animaï¿½ï¿½es faciais personalizadas em avatares. Master's thesis, Pontifi¿½cia Universidade Catï¿½lica do Rio Grande do Sul - PUCRS, Porto Alegre, RS, 2010.

[60] Rhee, T., Hwang, Y., Kim, J. D., and Kim, C. Real-time facial animation from live video tracking. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2011), SCA '11, ACM, pp. 215–224.

[61] Rumelhart, D. E., Hinon, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature 323* (October 1986), 533–536.

[62] Russell, J. A. Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies. *Psychological Bulletin* (1994), 102–141.

[63] Saragih, J. M., Lucey, S., and Cohn, J. F. Real-time avatar animation from a single image. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (march 2011), pp. 117–124.

[64] Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, B. Schouten, N. C. Juul, A. Drygajlo, and M. Tistarelli, Eds. Springer-Verlag, Berlin, Heidelberg, 2008, pp. 47–56.

[65] Schramm, R. Detecção de faces e rastreamento de pose da cabeï¿½a. Master's thesis, Universidade do Vale do Rio dos Sinos - UNISINOS, Sï¿½o Leopoldo, RS, 2009.

[66] Senechal, T., Rapp, V., Salam, H., Seguier, R., Bailly, K., and Prevost, L. Facial action recognition combining heterogeneous features via multikernel learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42*, 4 (Aug 2012), 993–1005.

[67] Seyedarabi, H., Lee, W., and Aghagolzadeh, A. Automatic lip tracking and action units classification using two-step active contours and probabilistic neural networks. In *Electrical and Computer Engineering, 2006. CCECE '06. Canadian Conference on* (May 2006), pp. 2021–2024.

[68] Sharifi, M., Fathy, M., and Tayefeh Mahmoudi, M. A classified and comparative study of edge detection algorithms. In *Information Technology: Coding and Computing, 2002. Proceedings. International Conference on* (2002), IEEE, pp. 117–120.

[69] Sheerman-Chase, T., Ong, E.-J., and Bowden, R. Non-linear predictors for facial feature tracking across pose and expression. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (April 2013), pp. 1–8.

[70] Sohail, A. S. M., and Bhattacharya, P. *Detection of Facial Feature Points Using Anthropometric Face Model*, vol. 31. Springer US, 2008, pp. 189–200.

[71] Stoiber, N., Seguier, R., and Breton, G. Facial animation retargeting and control based on a human appearance space. *Computer Animation and Virtual Worlds 21*, 1 (2010), 39–54.

[72] Strang, G. *Álgebra Linear e suas Aplicações*, 4 ed. Cengage Learning, 2009.

[73] Tang, H., and Huang, T. S. Mpeg4 performance-driven avatar via robust facial motion tracking. In *International Conference on Computer Vision* (2008), pp. 249–252.

[74] Taylor, G. W., and Hinton, G. E. Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th Annual International Conference on Machine Learning* (New York, NY, USA, 2009), ICML '09, ACM, pp. 1025–1032.

[75] Tian, Y.-L., Kanade, T., and Cohn, J. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 2 (February 2001), 97 – 115.

[76] Tissot, H. C., Camargo, L. C., and Pozo, A. T. Treinamento de redes neurais feedforward: comparativo dos algoritmos backpropagation e differential evolution. In *Brazilian Conference on Intelligent Systems* (2012).

[77] Tong, Y., Wang, Y., Zhu, Z., and Ji, Q. Facial feature tracking using a multi-state hierarchical shape model under varying face pose and facial expression. *Pattern Recognition, International Conference on 1* (2006), 283–286.

[78] Torresani, L., Hackney, P., and Bregler, C. Learning motion style synthesis from perceptual observations. In *NIPS* (2006), pp. 1393–1400.

[79] Tresadern, P., Ionita, M., and Cootes, T. Real-time facial feature tracking on a mobile device. *International Journal of Computer Vision 96*, 3 (2012), 280–289.

[80] van Kuilenburg, H., Wiering, M., and den Uyl, M. A model based method for automatic facial expression recognition. In *Machine Learning: ECML 2005*, J. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, Eds., vol. 3720 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, pp. 194–205.

[81] Vogler, C., and Goldenstein, S. Facial movement analysis in asl. *Univers. Access Inf. Soc. 6*, 4 (2008), 363–374.

[82] Wang, Y., Liu, Z.-Q., and Zhou, L.-Z. Key-styling: learning motion style for real-time synthesis of 3d animation. *Computer Animation and Virtual Worlds 17*, 3-4 (2006), 229–237.

[83] Weise, T., Bouaziz, S., Li, H., and Pauly, M. Realtime performance-based facial animation. In *ACM SIGGRAPH 2011 papers* (New York, NY, USA, 2011), ACM, pp. 77–87.

[84] Wu, C. H., and McLarty, J. W. *Methods in Computational Biology and Biochemestry: Neural Networks and Genome Informatics*. Elsevier, 2000.

[85] Wu, T., Butko, N., Ruvolo, P., Whitehill, J., Bartlett, M., and Movellan, J. R. Multilayer architectures for facial action unit recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42*, 4 (Aug 2012), 1027–1038.

[86] Wu, Y., Wang, Z., and Ji, Q. Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2013), CVPR '13, IEEE Computer Society, pp. 3452–3459.

[87] xiang Chai, J., Xiao, J., and Hodgins, J. Vision-based control of 3d facial animation. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2003)* (2003).

[88] Yang, P., Liu, Q., and Metaxas, D. N. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters 30*, 2 (2009), 132 – 139. Video-based Object and Event Analysis.

[89] Yuen, P., Lai, J. H., and Huang, Q. Y. Mouth state estimation in mobile computing environment. In *Sixth IEEE Int. Conf. on Automatic Face and Gesture Recognition* (2004), IEEE Computer Society, pp. 705–710.

[90] Zhang, Y., Ji, Q., Zhu, Z., and Yi, B. Dynamic facial expression analysis and synthesis with mpeg-4 facial animation parameters. *IEEE Trans. Circuits Syst. Video Techn. 18*, 10 (2008), 1383–1396.

[91] Zhu, Y., De la Torre, F., Cohn, J., and Zhang, Y.-J. Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *Affective Computing, IEEE Transactions on 2*, 2 (April 2011), 79–91.