INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Captura e Reprodução de Expressões Faciais

JOÃO MIGUEL DA SILVA FONSECA
Outubro de 2017

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Captura e Reprodução de Expressões Faciais

João Miguel da Silva Fonseca

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática, Área de Especialização em
Sistemas Gráficos e Multimédia

Orientador: Prof. Doutor João Paulo Pereira
Co-orientador: Prof. Doutor António Vieira de Castro

Porto, outubro 2017

Resumo

Na dobragem de produções audiovisuais só são traduzidas as falas, existindo uma discrepância entre o áudio e o que a personagem diz (movimentos da face, boca, e lábios), que resulta em falhas na compreensão da fala e em dobragens pouco realistas. Nos últimos anos, têm sido desenvolvidos métodos de captura e reprodução de movimentos faciais que capturam os movimentos faciais de um ator de dobragem e reproduzem esses movimentos na face de um ator previamente gravado.

Este documento contém a análise e avaliação do estado da arte de métodos de captura e de reprodução de movimentos faciais, e a descrição de uma solução de captura e reprodução em tempo-real, utilizando uma câmara normal, desenvolvida para tentar resolver os problemas existentes com as dobragens tradicionais.

A solução implementada foi avaliada através de questionários efetuados, demonstrando qualidade ainda inferior às dobragens tradicionais.

Palavras-chave: Captura e reprodução de movimentos faciais, modelos faciais 3D, dobragem.

Abstract

In the dubbing of audio-visual productions, only the lines are translated, and there is a discrepancy between the audio and what the character says (movements of the face, mouth, and lips), which results in poor speech comprehension and unrealistic dubs. In recent years, methods of capturing and reproducing facial movements have been developed that capture the facial movements of a dubbing actor and reproduce these movements in the face of a previously recorded actor.

This document contains the analysis and evaluation of the state of the art in methods of capture and reproduction of facial movements, and the description of a real-time capture and reproduction solution, using a normal camera, developed to address existing problems with traditional dubbing.

The implemented solution was evaluated through questionnaires, showing a quality that is still inferior to traditional dubbing.

Keywords: Capture and re-enactment of facial movements, 3D Morphable Model, Dubbing.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos

3D	Tridimensional
BFM	*Basel Face Model*, Modelo 3D facial da Universidade de Basel, Suíça
GPU	*Graphics Processing Unit*, Unidade de Processamento Gráfico
LBP	*Local Binary Patterns*, Padrões Binários Locais
PCA	*Principal Component Analysis*, Análise de Componentes Principais
RGB	*Red Green Blue*, Vermelho Verde Azul
RGB-D	D de *Depth*, Profundidade
SVD	*Singular Value Decomposition*, Decomposição em Valores Singulares

1 Introdução

Este capítulo contém uma breve introdução ao contexto da dissertação e ao problema que se pretende resolver. Contém também uma breve análise do valor criado pela solução resultante da dissertação.

1.1 Contexto

Em muitos países, as produções audiovisuais (filmes, séries televisivas, documentários, etc.) estrangeiras são traduzidas através da legendagem (texto traduzido a acompanhar a imagem) ou da dobragem (gravação de voz traduzida sobre a voz original). Na Europa, países como a Alemanha, França, Espanha e Itália têm uma grande tradição de dobragem, enquanto que países como Portugal, Grécia e os países nórdicos utilizam mais a legendagem. Esta divisão deve-se principalmente a fatores políticos, culturais e económicos, como leis que proibiam as dobragens, ou o custo superior da dobragem em relação à legendagem.

O mercado para produções audiovisuais dobradas é ainda um mercado grande, mas tem vindo a perder contra a legendagem. Existem várias razões para o declínio deste mercado, como a globalização e a preferência pela legendagem para aprender línguas novas, e também os problemas na qualidade das dobragens.

1.2 Problema

Na dobragem só são traduzidas as falas, existindo uma discrepância entre o áudio e o que a personagem diz (movimentos da face, boca, e lábios), que pode resultar em falhas na compreensão da fala (Sumby e Pollack 1954). O áudio dobrado é também frequentemente misturado incorretamente com o áudio da cena.

Estes dois problemas resultam em dobragens de baixa qualidade e pouco realistas, e fazem parte das razões para o declínio do mercado das dobragens.

1.3 Objetivo

Para eliminar a discrepância entre as partes visuais e auditivas, é necessário modificar a parte visual (face da personagem) para que esteja sincronizada com a fala traduzida. Uma forma de fazer essa sincronização é a captura e reprodução (ou transferência) de movimentos faciais, onde os movimentos faciais feitos pelo ator tradutor são transferidos para a face da personagem.

Nos últimos anos, os métodos de captura e reprodução de movimentos faciais têm sido muito desenvolvidos, já não sendo preciso colocar marcadores/pontos de referência físicos na face. Recentemente, foi desenvolvido um sistema que utiliza uma câmara RGB (*Red Green Blue*, Vermelho Verde Azul) normal (sem profundidade) para capturar o movimento da cara de um ator, reproduzindo esse movimento na cara de um ator alvo (gravada em tempo-real ou de um vídeo previamente gravado).

No entanto, estes sistemas ainda sofrem de muitos problemas, nomeadamente a presença de artefactos na cara do ator no resultado final, como a deformação dos lábios. Para reduzir a discrepância entre as partes visuais e auditivas, e melhorar a experiência de visualização por parte dos espetadores, é necessário aperfeiçoar estes sistemas.

Assim, o objetivo principal desta dissertação é desenvolver uma solução de captura e reprodução de movimentos faciais e, se possível, tentar resolver alguns desses problemas, de forma a tornar as dobragens mais realistas.

1.4 Resultados Esperados

No final desta dissertação espera-se que seja possível melhorar as dobragens de produções audiovisuais através da captura e reprodução de movimentos faciais, e que se tenha resolvido alguns dos problemas e limitações dos sistemas existentes.

1.5 Análise de Valor

A solução desenvolvida permitirá às empresas de dobragem sincronizar os movimentos faciais de um ator com o áudio dobrado, aumentando o realismo da dobragem e resultando numa melhor experiência de visualização. Embora haja um ligeiro aumento no custo da produção da dobragem, já que terá de ser comprada a solução e o material necessário à sua utilização, o aumento da satisfação dos espetadores (e até do número de espetadores) deverá compensar os novos custos.

1.6 Abordagem Preconizada

A abordagem preconizada inclui o design e desenvolvimento de uma aplicação de *desktop* de captura e reprodução de movimentos faciais, que permita a visualização dos resultados em tempo real, ou seja, enquanto o ator faz a dobragem, e que utilize uma câmara RGB normal. O sistema será comparado com sistemas existentes, para se verificar se são obtidos resultados melhores, e as dobragens criadas pela aplicação serão validadas através de questionários.

1.7 Estrutura do Documento

O resto deste documento está dividido em cinco capítulos. No capítulo 2, Contexto e Estado de Arte, é descrito em mais detalhe o contexto desta dissertação, juntamente com a análise de valor segundo o respetivo módulo curricular, e é também apresentado o estado da arte. O capítulo 3 avalia o estado da arte, e contém a escolha (e a sua justificação) do método base escolhido para a solução, juntamente com uma secção para apresentar as limitações desse método. O capítulo 4 apresenta o *design* da solução desenvolvida, e o capítulo 5 descreve detalhadamente os algoritmos implementados. O capítulo 6, Avaliação da Solução, descreve a forma como a solução desenvolvida foi avaliada e comparada a soluções existentes. O capítulo 7 contém a conclusão do trabalho efetuado e uma discussão sobre trabalho futuro.

2 Contexto e Estado da Arte

Este capítulo contém uma apresentação detalhada do contexto da dissertação, uma análise de valor segundo o respetivo módulo curricular e uma apresentação do estado da arte.

2.1 Contexto

Em muitos países, as produções audiovisuais (filmes, séries televisivas, documentários, etc.) estrangeiras são traduzidas através da legendagem (texto traduzido a acompanhar a imagem) ou da dobragem (gravação de voz traduzida sobre a voz original). Na Europa, países como a Alemanha, França, Espanha, e Itália têm uma grande tradição de dobragem, enquanto que países como Portugal, Grécia, e os países nórdicos utilizam mais a legendagem, deixando a dobragem para as produções animadas e para as produções direcionadas para as crianças. Alguns países, como a Polónia, mantêm a voz original, e colocam por cima uma voz traduzida.

Esta divisão em dois grupos (legendagem e dobragem) deve-se a fatores políticos, culturais e económicos. Nos fatores políticos temos, por exemplo, Portugal e Espanha. Portugal, em 1948, passou uma lei que bania a dobragem, numa tentativa de proteger a indústria nacional (Presidência do Conselho 1948). Isto levou a uma falta de empresas interessadas em fazer dobragem, mesmo depois da lei desaparecer. Em Espanha, devido à censura imposta pelo governo, todos as produções estrangeiras eram dobradas (Higginbotham 1988). Nos fatores culturais, temos o resultado da tradição: quem está habituado a ver produções dobradas não aceita ver produções legendadas, e vice-versa. Nos fatores económicos encontra-se a maior diferença entre os dois métodos, já que a dobragem é muito mais cara que a legendagem (tradução, estúdio, equipamento, atores versus apenas tradução). Assim, os países com mais população conseguiam suportar os custos superiores, mas os países com menos população optavam pela legendagem.

Num estudo feito em 2012 para a Comissão Europeia (European Commission 2012), 52% dos europeus preferiam dobragens em vez de legendagens (44%). Existe, portanto, um mercado grande para as produções dobradas. No entanto, este mercado está a diminuir. No mesmo estudo, feito em 2006 (European Commission 2006), 56% preferiam dobragens, e 37%

preferiam legendagens. No estudo de 2012, 55% dos europeus de idades entre os 14 e os 24 anos preferiam legendagens, mas só 35% acima dos 55 anos preferiam o mesmo. À medida que as pessoas adoptam mais línguas (nomeadamente o Inglês), há uma tendência a preferirem a legendagem.

2.2 Problema

Um dos problemas que leva muitas pessoas a não gostarem das dobragens é a sua baixa qualidade. O áudio dobrado é frequentemente misturado incorretamente com o áudio da cena, e existe uma discrepância entre o áudio e o que a personagem diz (movimentos da face, boca, e lábios). Estes problemas levam a uma falta de realismo e a uma sensação de anormalidade (Koolstra et al. 2002). Inclusivamente, a diferença entre as falas e os movimentos faciais pode resultar em falhas na compreensão da fala (Sumby e Pollack 1954). O efeito McGurk (McGurk e MacDonald 1976), por exemplo, demonstra a relação entre a audição e a visão na perceção da fala, e ocorre quando uma pessoa ouve um determinado som e vê ao mesmo tempo os movimentos faciais de outro som, mas acaba por ouvir um terceiro som.

A falta de realismo e a sensação de anormalidade podem funcionar como uma barreira que impede novos espetadores de aderirem às dobragens. Portanto, é necessário encontrar soluções para estes problemas. Uma solução que tem vindo a ser investigada é a sincronização entre o áudio e a parte visual, através do controlo dos movimentos faciais das personagens. Existem várias formas de controlar os movimentos faciais, como o controlo a partir do áudio dobrado ou através da captura e reprodução de movimentos faciais do ator que faz a dobragem. No entanto, estes sistemas ainda sofrem de muitos problemas, nomeadamente a presença de artefactos na cara do ator alvo, como a deformação dos lábios.

O propósito deste projecto é desenvolver uma solução de captura e reprodução de movimentos faciais, avaliar as dobragens obtidas pela solução e, se possível, resolver alguns dos problemas das soluções existentes, de forma a tornar os resultados mais realistas, diminuir a discrepância entre as partes visuais e auditivas e melhorar a experiência de visualização por parte dos espetadores.

Como foi referido anteriormente, a dobragem é mais dispendiosa do que a legendagem. A legendagem resume-se à tradução e colocação do texto no ecrã, mas a dobragem necessita de tradução, estúdios de gravação, atores de voz profissionais, técnicos de som, entre outros. Para tornar a solução desenvolvida mais atrativa para as empresas de dobragem, é preciso reduzir ao máximo o custo envolvido na sua utilização. Embora algumas técnicas relacionadas com a captura de movimentos faciais utilizem marcadores faciais, ou várias câmaras colocadas à volta do ator, a tendência tem sido em desenvolver métodos que utilizem câmaras normais RGB. A solução desenvolvida também deverá funcionar em tempo real, de forma a não aumentar o tempo gasto na dobragem.

2.3 Análise de Valor

Nesta secção é feita uma análise do valor da solução resultante desta dissertação, segundo o módulo curricular. Nas subsecções seguintes é feita uma descrição do modelo de Peter Koen de desenvolvimento de novos conceitos, uma descrição dos benefícios e sacrifícios resultantes da utilização da solução a desenvolver, e é enunciada a proposta de valor dessa solução. A ultima subsecção contém o modelo Canvas, que é utilizado para descrever uma possível ideia de negócio que utilizasse a solução como seu produto.

2.3.1 Modelo de Desenvolvimento de Novos Conceitos

O modelo de desenvolvimente de novos conceitos (Koen et al. 2001) define componentes para ajudar na definição de oportunidades, ideias, e conceitos, antes de se passar para o desenvolvimento de um produto ou serviço. Contém cinco elementos-chave: identificação de oportunidades, análise de oportunidades, geração de ideias, seleção de ideias, e definição de conceitos. Nesta subsecção, são descritos estes elementos-chave, alguns métodos e ferramentas utilizadas em cada um, e é feita a contextualização desses elementos no contexto desta dissertação. Embora esta secção siga a ordem dos elementos-chave, o modelo, como se pode ver na Figura 1, permite e incentiva a iteração entre os elementos, podendo ir de uma ideia para a identificação de oportunidades.

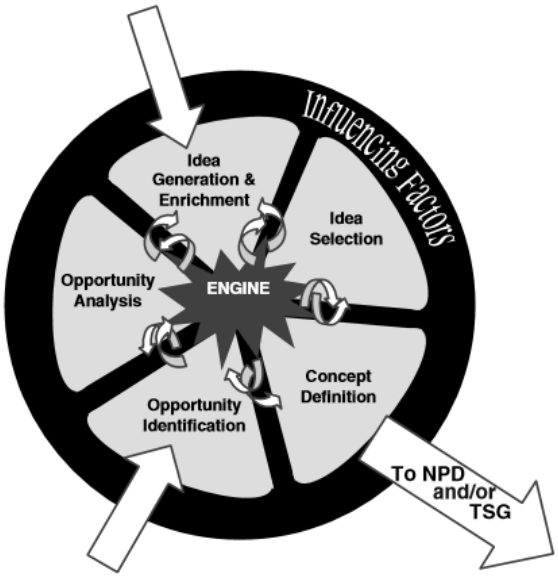

Figura 1: O modelo de desenvolvimento de novos conceitos.
Fonte: (Koen et al. 2002)

Identificação de oportunidades é quando uma organização identifica oportunidades que possam ser do seu interesse, como uma falha nas tecnologias existentes, um novo processo que reduza os custos, entre outros. Estas oportunidades são normalmente identificadas através da análise dos mercados, das tecnologias existentes ou em desenvolvimento, e dos interesses dos clientes. Neste projecto, foi identificada uma oportunidade no mercado da tradução e dobragem de produções audiovisuais, onde existe um problema de sincronização entre os movimentos faciais do ator e o áudio traduzido.

Após identificadas as oportunidades, elas têm de ser analisadas para se verificar se são realmente boas oportunidades. As oportunidades são analisadas com os mesmos métodos usados para as identificar, mas com um maior foco em encontrar justificações para a escolha de uma oportunidade. A oportunidade identificada insere-se num mercado grande, e vários países europeus têm uma grande tradição de dobragem, mas a falta de sincronização leva a falhas na compreensão e a uma sensação de anormalidade (Koolstra et al. 2002) que pode resultar numa redução dos espetadores das produções dobradas.

Após ser escolhida uma oportunidade, é necessário gerar ideias que satisfaçam os problemas existentes nessa oportunidade. As ideias passam por várias fases de construção, discussão, e desenvolvimento, até se chegar a uma ideia concreta. As ideias aparecem através de sessões de brainstorming, incentivos que estimulem a proposta de ideias, etc. Para resolver o problema da sincronização existem várias alternativas a serem investigadas, como o controlo dos movimentos faciais através do áudio dobrado, ou a captura e reprodução de movimentos faciais de outro ator.

As ideias são depois selecionadas tendo em conta os riscos e os benefícios resultantes da sua escolha. Além da opinião e experiência pessoal das pessoas que façam a seleção, também são utilizados métodos de análise de risco, probabilidade de sucesso comercial, entre outros. O sistema base para este projecto foi escolhido pelos resultados recentes (qualidade, execução em tempo real, e uso de uma câmara RGB) que demonstram um risco reduzido e uma grande probabilidade de sucesso à medida que a tecnologia for evoluindo.

O elemento-chave final é a definição do conceito, onde se define o que é necessário para a solução pretendida, como os objetivos, mercado, e necessidades dos clientes, com a finalidade de ter um documento que justifique o investimento necessário e a passagem para o desenvolvimento da solução. Para definir e avaliar os conceitos é necessária uma análise rigorosa de todos os fatores envolvidos (tamanho do mercado, tecnologia existente, potencial do negócio, etc.). Neste projecto, o conceito final seria uma aplicação que fizesse a captura e reprodução dos movimentos faciais do ator da dobragem para o ator original, e que se pudesse integrar sem dificuldades no processo de dobragem.

2.3.2 Benefícios e Sacrifícios

Esta subsecção define três conceitos relativos à análise do valor de um produto ou serviço, necessários para a definição dos benefícios e sacrifícios para o cliente que adquira a solução.

Ao analisarmos o valor do produto, podemos ter em conta três conceitos-chave: valor (*value*), valor percebido (*perceived value*), e valor para o cliente (*value for the customer*).

O valor de um produto ou serviço, em relação aos outros produtos, é o que este adiciona a um negócio (resultando no aumento dos lucros e na descida dos custos), ou dá a um consumidor (resultando, por exemplo, na melhoria da qualidade de vida). Um produto adiciona valor a um negócio ao, por exemplo, ser de maior qualidade, ter mais funcionalidades úteis, e ser mais eficaz e eficiente. A um consumidor, um produto adiciona valor se, por exemplo, tiver uma maior usabilidade e tiver um custo reduzido. A solução a desenvolver adiciona valor a um negócio (empresas de dobragem) na forma de um aumento na qualidade das produções dobradas por esse negócio, o que leva ao aumento dos seus lucros.

Valor percebido é o valor que um cliente atribui a um produto tendo em conta os benefícios e os sacrifícios que advêm da compra e utilização do produto. Os benefícios incluem a qualidade superior do produto e a capacidade de customização do produto à situação específica do cliente, etc. Os sacrifícios consistem principalmente no custo monetário do produto e como se compara ao custo dos produtos dos competidores.

Valor para o cliente é a perceção das vantagens que o cliente tem ao utilizar o produto, tendo em conta se o produto reduz sacrifícios, se beneficia o cliente, e como o custo se compara ao custo dos produtos competidores. Woodall (2003) divide o valor para o cliente em quatro fases de uma perspetiva longitudinal (ou seja, ao longo do tempo): antes da compra, durante a compra, após a compra, e após o uso do produto.

Tendo definido estes três conceitos, passamos à definição dos benefícios e sacrifícios.

Podemos dividir os benefícios em dois grupos: os benefícios que um sistema de captura e reprodução de movimentos faciais tem em relação à dobragem tradicional, e os benefícios que o resultado final desta dissertação terá em relação aos sistemas já existentes. Em relação à dobragem tradicional, estes sistemas resultam numa dobragem de maior qualidade ao sincronizar a parte visual com a parte auditiva. A qualidade superior resultará num aumento dos lucros das empresas de dobragem, já que aumenta a satisfação dos espetadores. Em relação aos sistemas já existentes, o resultado final desta dissertação será um sistema com resultados mais realistas, o que resulta numa dobragem de maior qualidade, mas em tempo útil, ao contrário de alguns sistemas. Outro benefício será a constante melhoria da solução ao longo dos anos.

Os sacrifícios consistem num aumento dos custos e do tempo de produção. Para utilizar um sistema destes será preciso gastar mais dinheiro, tanto na compra do sistema, como na compra de uma câmara para gravar o ator. Será também necessário gastar mais tempo na produção da dobragem, para maximizar a qualidade da dobragem.

Colocando numa perspetiva longitudinal, antes da compra do produto, o cliente (empresa de dobragem) tem as dobragens tradicionais e acredita que o produto poderá criar valor, já que resultará em dobragens de maior qualidade. Após a compra, o cliente tem a perceção das vantagens que tem ao utilizar o produto, tendo em conta os benefícios (dobragens de maior qualidade) e os sacrifícios (produções mais demoradas).

2.3.3 Proposta de Valor

A proposta de valor define qual é o produto, quem são os clientes, e que valor o produto cria. Assim, a proposta de valor da solução a desenvolver pode ser definida da seguinte forma.

Este produto é um sistema de captura e reprodução de movimentos faciais que permite às empresas de dobragem sincronizar os movimentos faciais do ator com o áudio dobrado, melhorando assim a qualidade da dobragem. A execução em tempo real e com uma câmara RGB normal permite ver os resultados de forma rápida e com uma configuração mínima.

2.3.4 Modelo Canvas

O modelo Canvas (Figura 2) descreve uma possível ideia de negócio que utiliza como seu produto a solução a desenvolver nesta dissertação.

<table>
<tr>
<td rowspan="2">Parcerias-chave
Apoio a startups do ISEP.</td>
<td>Actividades-chave
Investigação e desenvolvimento para melhorar continuamente o produto;

Serviço de apoio ao cliente na utilização do produto.</td>
<td rowspan="2">Proposta de valor
Sistema de captura e reprodução de movimentos faciais que permite às empresas de dobragem sincronizar os movimentos faciais do ator com o áudio dobrado, melhorando assim a qualidade da dobragem;

Execução em tempo real e com uma câmara RGB normal permite ver os resultados de forma rápida e com uma configuração mínima.</td>
<td>Relacionamento com Clientes
Relação próxima, com assistência na utilização, e colaboração para melhorar o produto.</td>
<td rowspan="2">Segmentos de clientes
Empresas de tradução e dobragem de produções audiovisuais (filmes/séries televisivas/ documentários, etc.).</td>
</tr>
<tr>
<td>Recursos-chave
Equipas de investigação, desenvolvimento, e manutenção (apoio ao cliente) do produto.</td>
<td>Canais
E-mail;

Plataforma online.</td>
</tr>
<tr>
<td colspan="3">Estrutura de custos
Salários dos funcionários;
Desenvolvimento e manutenção do produto.</td>
<td colspan="2">Fontes de Receita
Licença de utilização do produto.</td>
</tr>
</table>

Figura 2: Modelo de negócio Canvas

Os clientes são as empresas de tradução e dobragem de produções audiovisuais, com as quais será mantida uma relação próxima de forma a poder prestar assistência na utilização do produto. Esta relação próxima também resulta num *feedback* constante, que pode ser utilizado para melhorar o produto.

Haverá uma constante melhoria do produto, através de investigação e desenvolvimento. Isto leva a que os recursos-chave sejam as equipas de investigação e desenvolvimento, além das equipas de manutenção.

Para lançar o produto e ajudar em todo o processo, será constituída uma parceria com o ISEP.

A fonte de receita será a licença que as empresas de dobragem terão de pagar para adquirir o produto, e os custos estarão maioritariamente nos salários das equipas e no desenvolvimento e manutenção do produto.

2.4 Estado da Arte

Esta secção apresenta o estado da arte da captura e reprodução de movimentos faciais, e apresenta também algumas tecnologias relevantes.

2.4.1 Captura de Movimentos Faciais

A captura dos movimentos faciais é a componente mais importante nos métodos de captura e reprodução. Há muitos anos que é usada na produção de filmes e jogos, utilizando marcadores faciais (Huang et al. 2011), múltiplas câmaras (*camera arrays*) colocadas em redor do ator, etc. Embora os resultados sejam de alta qualidade, são métodos caros, que precisam de iluminação controlada, e pouco confortáveis para os atores, por isso a tendência tem sido de ir na direção de métodos sem marcadores e com menos câmaras.

Recentemente têm aparecido métodos de captura utilizando apenas uma câmara. Alguns são métodos *offline*, ou seja, que não correm em tempo real e utilizam toda a informação do vídeo (Shi et al. 2014; Garrido et al. 2013; Garrido, Zollhöfer, Casas, et al. 2016). Outros são *online*, e não precisam de ter o vídeo todo (Cao, Hou, et al. 2014; Thies et al. 2016).

Os movimentos faciais são normalmente modelados utilizando *blendshapes* ou modelos multilineares. As *blendshapes* são muito utilizadas na indústria da animação, e consistem em modelos 3D da expressão neutra da face e de várias outras expressões (frequentemente é guardada só a diferença de cada expressão à expressão neutra), e uma expressão é obtida através de uma combinação linear de algumas expressões. Os modelos multilineares descrevem um modelo 3D de uma face não só através das expressões, mas também através da identidade, textura, iluminação, e outros fatores, dependendo do que se pretende capturar.

É possível gerar um modelo da face a partir de uma imagem através da otimização (Shi et al. 2014; Thies et al. 2016; Garrido, Zollhöfer, Casas, et al. 2016) ou regressão (Cao et al. 2013; Cao, Hou, et al. 2014).

Na otimização, tenta-se ajustar um modelo genérico à imagem, de forma a obter o modelo 3D que representa a face presente na imagem. São utilizados detectores de pontos faciais para ajustar a geometria (posição da boca, olhos, nariz, etc.) do modelo aos pontos detectados, e são também tidos em consideração outros fatores, como a diferença entre duas imagens sucessivas (que tem de ser minimizada), ou a diferença entre a imagem sintetizada e a imagem real (Thies et al. 2016). O modelo genérico é criado a partir de bases de dados de faces capturadas com métodos de alta qualidade (Cao, Weng, et al. 2014).

Na regressão, é treinado um regressor com imagens e modelos 3D de alta qualidade das faces dessas imagens, e durante a execução o regressor infere o modelo 3D referente à imagem dada como input. Cao et al. (2013) treinam um regressor específico para cada pessoa, e Cao, Hou, et al. (2014) treinam um regressor geral que funciona com pessoas que não fizeram parte do treino.

Os métodos de uma só câmara não conseguem capturar os pequenos detalhes da face, como as rugas (Cao, Hou, et al. 2014). Para obter esses detalhes após a captura dos movimentos faciais, pode-se, mais uma vez, utilizar a otimização (Garrido et al. 2013; Shi et al. 2014) ou a regressão (Cao et al. 2015; Garrido, Zollhöfer, Casas, et al. 2016). Na otimização, sintetiza-se uma imagem, simulando a iluminação da cena real, e tenta-se minimizar a diferença entre a imagem sintetizada e a imagem real. Na regressão, treina-se um regressor com modelos 3D de alta qualidade e os respetivos modelos sem detalhes, e utiliza-se o regressor para, dado um modelo sem detalhe, obter os detalhes que faltam. Por exemplo, o método de (Cao, Hou, et al. 2014) é melhorado com um regressor em (Cao et al. 2015).

A Figura 3 demonstra a captura de movimentos faciais e adição de detalhes (Shi et al. 2014), ambas através da otimização.

Figura 3: Exemplo de captura de movimentos e da adição de detalhes. Da esquerda para a direita: imagem de *input*, modelo capturado, modelo sobreposto à imagem, e modelo com detalhes.
Fonte: (Shi et al. 2014)

2.4.2 Controlo dos Movimentos Faciais

Têm sido desenvolvidas várias tentativas de controlar os movimentos da face de um ator. O método descrito por Bregler et al. (1997) cria uma base de dados com as imagens da boca do ator referentes aos diferentes fonemas, e reordena as imagens de acordo com o novo áudio. O método de Malleson et al. (2016) permite misturar movimentos faciais de um ator, mas está limitado aos movimentos previamente gravados, tal como o de Bregler et al. (1997). Garrido et al. (2015) capturam os movimentos faciais do ator da dobragem e transferem-nos para a face do ator original, e utilizam o áudio traduzido para melhorar a sincronização entre o áudio e os lábios. Este método foi desenvolvido especificamente para tradução e dobragem, mas não funciona em tempo real e necessita do vídeo completo do ator da dobragem.

A captura e reprodução de movimentos faciais em tempo real foi demonstrada por Thies et al. (2015), utilizando uma câmara com informação de profundidade (RGB-D (D de *Depth*, Profundidade)). Neste método, combinam um modelo de identidade e reflexão da pele com um modelo das expressões faciais. Os movimentos são depois transferidos do modelo do ator da dobragem para o modelo do ator original. Este método foi melhorado posteriormente de forma a funcionar com uma câmara RGB normal (Thies et al. 2016). Um resultado deste método pode ser visto na Figura 4. Mesmo sem informação sobre o áudio, os autores demonstraram resultados comparáveis aos de Garrido et al. (2015), mas em tempo real. A execução em tempo real é conseguida através da utilização da Unidade de Processamento Gráfico (GPU, *Graphics Processing Unit*) para encontrar uma solução para o algoritmo de Gauss-Newton, que é utilizado na captura de movimentos.

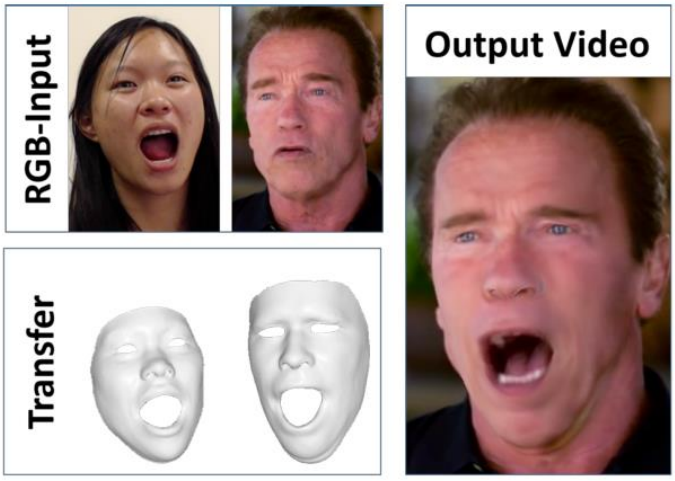

Figura 4: Exemplo da captura e reprodução de movimentos faciais.
Fonte: (Thies et al. 2016)

2.4.3 Síntese do Interior da Boca

Os métodos de captura não conseguem capturar o interior da boca (dentes e língua), devido a oclusões e à falta de luz. Para sintetizar o interior da boca ao controlar os movimentos da face, alguns métodos copiam a boca directamente do vídeo original (Vlasic et al. 2005) ou usam um modelo 3D falso (Garrido et al. 2015). Em (Thies et al. 2016) é criado um grafo de todas as imagens da boca durante o vídeo, e é usada a imagem da boca mais adequada aos novos movimentos. Recentemente, foi desenvolvido um método que obtém um modelo 3D dos dentes, utilizando um modelo obtido a partir de digitalizações 3D de modelos em gesso de dentes (Wu et al. 2016).

2.4.4 Tecnologia Relevante

A captura de movimentos faciais utiliza pontos faciais (*landmarks*) para reduzir os erros na criação do modelo da cara. Estes pontos faciais incluem os contornos dos olhos, da boca e da face, e o nariz, entre outros. Como o detetor de pontos faciais é um componente pequeno do algoritmo, não será feito o estado da arte, mas existem estudos que analisam e comparam exaustivamente os vários detectores existentes (Jin e Tan 2016; Chrysos et al. 2017). Existem várias bibliotecas que fornecem implementações de detectores de pontos faciais, como OpenFace[1] (Baltrusaitis et al. 2016), FaceTracker[2], e Dlib[3].

2.5 Conclusão

Neste capítulo foi detalhado o contexto da dissertação e o problema que se pretende resolver. Foi efetuada uma análise de valor da solução resultante, foi apresentado o estado da arte de métodos de captura e reprodução de movimentos faciais, e foram mencionadas algumas tecnologias relevantes a esses métodos.

[1] https://github.com/TadasBaltrusaitis/OpenFace
[2] https://github.com/kylemcdonald/FaceTracker
[3] http://dlib.net/

3 Avaliação do Estado da Arte

Neste capítulo são avaliados alguns dos métodos enunciados no estado da arte. Contém a escolha (e a sua justificação) do método base escolhido para a solução a desenvolver, e uma secção para apresentar as limitações desse método.

Convém mencionar que muitos dos artigos avaliados (e outros dentro da mesma área) não apresentam os resultados com valores numéricos que representem a qualidade do seu método em relação aos outros, limitando-se somente a comparar os resultados visualmente (modelo 3D colocado sobre a face). Isto talvez se deva à falta de métodos de comparação, ou à falta de consenso sobre qual é o objetivo final que os métodos querem atingir. Assim, a avaliação feita neste capítulo utiliza principalmente a componente visual.

3.1 Captura e Reprodução de Movimentos Faciais

Na avaliação dos métodos de reprodução/transferência de movimentos faciais, tem de se ter em conta dois fatores: a qualidade do resultado, e se é ou não executável em tempo real.

Quanto à qualidade, a Figura 5 e a Figura 6 comparam três métodos (Garrido et al. 2015; Thies et al. 2015; Thies et al. 2016). Em imagens retiradas dos resultados finais, os três métodos demonstram qualidade similar, sendo que a maior diferença está no interior da boca. Em vídeos[4, 5], o método de Thies et al. (2016) demonstra maior realismo nos movimentos, principalmente quando comparado com o de Garrido et al. (2015).

[4] https://www.youtube.com/watch?v=ohmajJTcpNk
[5] http://gvv.mpi-inf.mpg.de/projects/VisualDubbing/

Figura 5: Comparação entre (Garrido et al. 2015) e (Thies et al. 2016).
Fonte: (Thies et al. 2016)

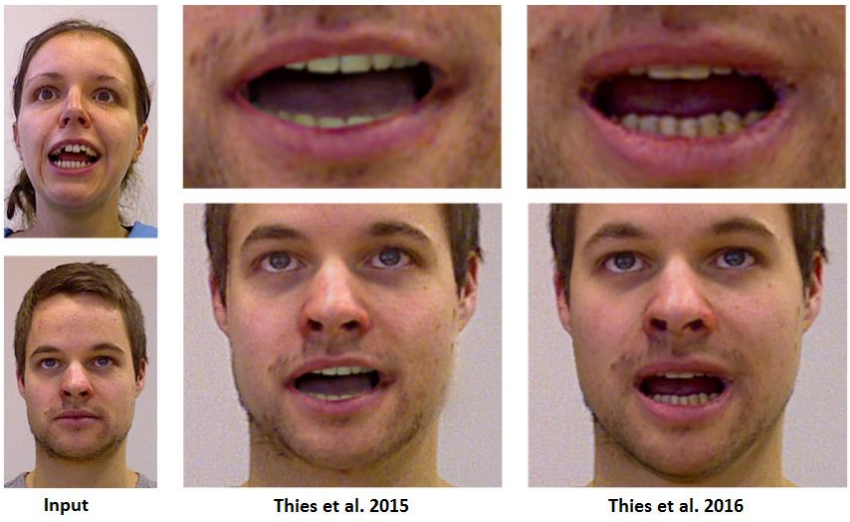

Figura 6: Comparação entre (Thies et al. 2015) e (Thies et al. 2016).
Fonte: (Thies et al. 2016)

Os autores de (Garrido et al. 2015) fizeram um inquérito em que mostravam três pares de vídeos, e cada par consistia num vídeo com dobragem tradicional e outro com a dobragem feita pelo método deles. Em todos os pares, as dobragens tradicionais conseguiram pontuações superiores, e 65% das pessoas preferiram as dobragens tradicionais às novas.

Quanto ao tempo de execução, como mencionado previamente, os métodos de Thies et al. (2015) e Thies et al. (2016) são os únicos que funcionam em tempo real. Segundo os resultados disponibilizados, o método de Thies et al. (2015) consegue processar 30 imagens por segundo, e o método de Thies et al. (2016) consegue processar 28 imagens por segundo.

3.2 Captura de Movimentos Faciais

Os métodos de captura de movimentos faciais também podem ser avaliados com os mesmos fatores.

Quanto à execução em tempo real, os métodos de (Shi et al. 2014; Garrido et al. 2013; Garrido, Zollhöfer, Casas, et al. 2016) são métodos *offline*. Por exemplo, o método de Garrido, Zollhöfer, Casas, et al. (2016) demora mais de 15 segundos a processar uma imagem. Outros métodos são *online*, como (Cao, Hou, et al. 2014; Thies et al. 2016). Estes dois métodos conseguem ambos processar, em média, 28 imagens por segundo.

Quanto à qualidade da captura, Garrido, Zollhöfer, Casas, et al. (2016) apresentam um dos melhores resultados até agora, como se pode ver na Figura 7.

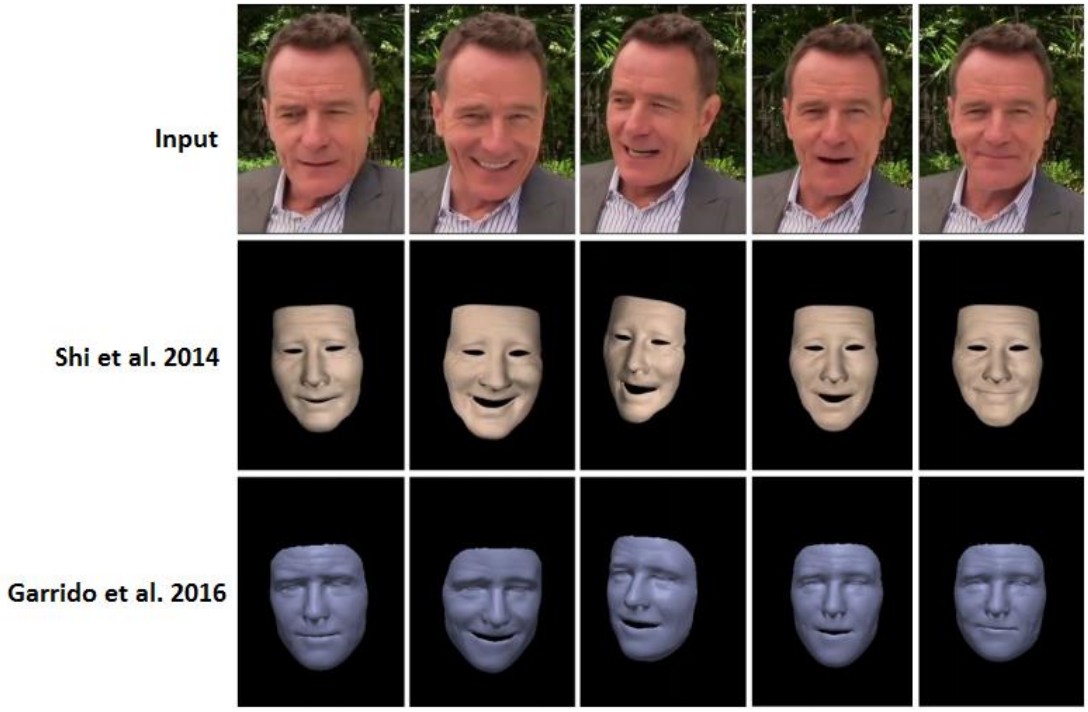

Figura 7: Comparação entre (Shi et al. 2014) e (Garrido, Zollhöfer, Casas, et al. 2016).
Fonte: (Garrido, Zollhöfer, Casas, et al. 2016)

Os autores de (Garrido, Zollhöfer, Casas, et al. 2016) mencionam que o método deles poderia fazer uso da GPU, tal como (Thies et al. 2016), o que levaria a um aumento na velocidade de processamento. Caso isto aconteça, o método poderá, no futuro, ser uma boa opção para integração com os sistemas de captura e reprodução de movimentos faciais.

Segundo uma comparação feita com vídeos recolhidos por Valgaerts et al. (2012), o método de Garrido, Zollhöfer, Casas, et al. (2016) atingiu um erro geométrico médio (distância euclidiana entre pontos dos modelos 3D) menor que 1 milímetro. O método de Thies et al.

(2015) conseguiu um erro de 1.5 milímetros, com um máximo de 7.9 milímetros. Pelos resultados demonstrados por Thies et al. (2016), os métodos de Cao, Hou, et al. (2014), Thies et al. (2015), e Thies et al. (2016) têm qualidade similar.

3.3 Escolha do Método a Usar

Tendo em conta o que foi descrito até agora, a melhor escolha em termos de qualidade e velocidade de processamento é o método de Thies et al. (2016). Este será, por isso, o método base desta dissertação. No entanto, convém mencionar algumas das limitações, não só as que foram mencionadas pelos autores, mas também as que são visíveis nos vídeos de demonstração do método.

3.4 Limitações

Existem seis limitações principais: os autores mencionam quatro limitações no artigo, e foram identificadas duas nos vídeos de demonstração.

O algoritmo assume que a pele é uma superfície Lambertiana e que a luz incide da mesma maneira em toda a superfície. Esta é uma limitação partilhada por muitos métodos de captura de movimentos, e pode levar a erros na captura caso existam sombras ou áreas brilhantes na face.

Outra limitação está relacionada com a oclusão da cara, por exemplo por cabelo, barba, mãos, e óculos, que também resulta em erros na captura e impossibilita a reprodução dos movimentos nas imagens com oclusão.

A terceira limitação já foi mencionada previamente, e tem a ver com a falha de detalhes no modelo 3D da face, devido à utilização de uma só câmara. Esta falta de detalhes pode levar a resultados pouco realista, como uma pele demasiado lisa.

A quarta limitação resulta do método utilizado para sintetizar o interior da boca, que cria um grafo de todas as imagens da boca durante o vídeo, e utiliza a imagem da boca mais adequada aos novos movimentos. Caso não haja variação suficiente nos movimentos do ator original, não é possível sintetizar correctamente o interior da boca. Isto pode acontecer, por exemplo, se o ator da dobragem fizer um sorriso exagerado (boca muito aberta), mas o ator original não o tiver feito durante o vídeo. O resultado será um interior da boca deformado, já que não existe informação suficiente para sintetizar correctamente.

Uma das limitações identificadas no vídeo de demonstração está relacionada com a transferência de movimentos dos lábios, que por vezes falha e resulta em lábios posicionados incorretamente (Figura 8). Isto acontece por causa da falta de variação no vídeo original, e

também por causa de se utilizar só uma câmara, que não fornece informação suficiente sobre a estrutura 3D dos lábios.

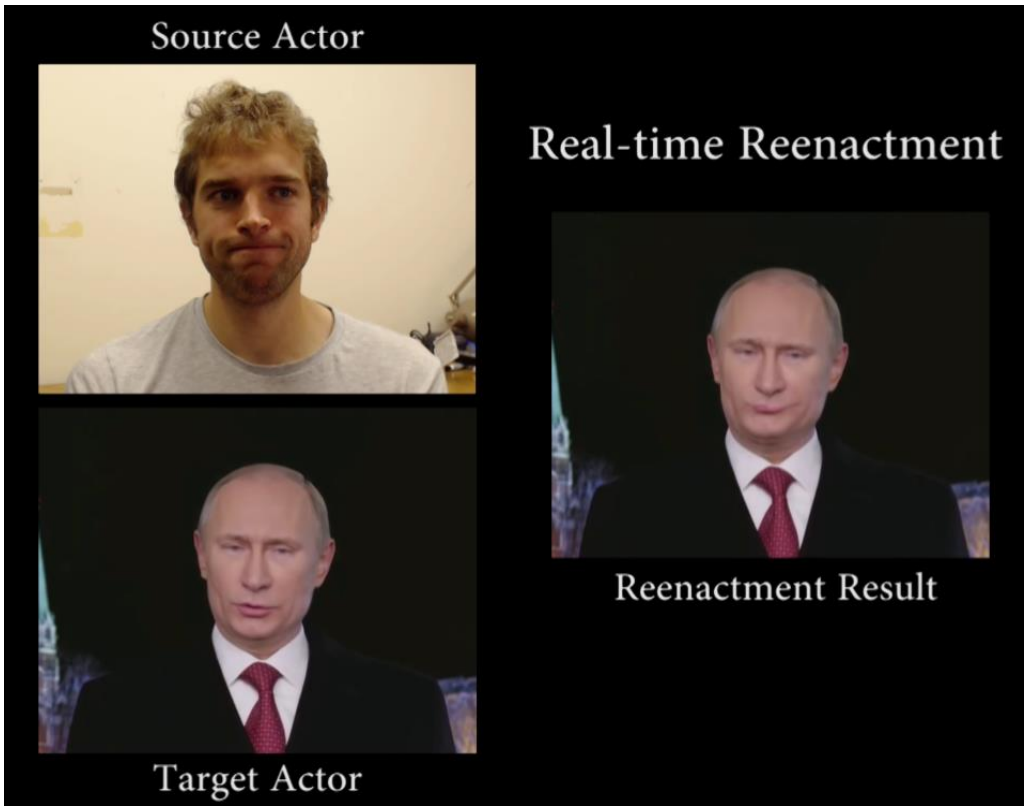

Figura 8: Exemplo de uma falha na reprodução de movimentos dos lábios.
Fonte: (Matthias Niessner 2016)

Outra limitação identificada é a falta de controlo sobre os olhos. Este método não captura os movimentos dos olhos, e não reproduz o movimento no ator. Dependendo dos movimentos faciais, os olhos podem ser uma parte crucial para tornar a dobragem mais realista. É por isso importante conseguir capturar e reproduzir o movimento dos olhos.

Existem várias soluções possíveis para estas limitações. Para adicionar detalhes à cara, pode-se treinar um regressor com modelos de alta qualidade (Cao et al. 2015). Para controlar os olhos, pode-se criar um modelo 3D dos olhos (Wood et al. 2016). Em (Wu et al. 2016) é criado um modelo 3D dos dentes, o que poderá melhorar a síntese do interior da boca. Recentemente, em (Garrido, Zollhöfer, Wu, et al. 2016), foi criada uma base de dados de modelos 3D de alta qualidade de lábios e utilizado um regressor para melhorar a captura dos lábios efectuada pelo método de Garrido, Zollhöfer, Casas, et al. (2016).

3.5 Deteção de pontos faciais

Segundo os estudos efectuados (Jin e Tan 2016; Chrysos et al. 2017), o método de Kazemi e Sullivan (2014) é dos métodos com melhores resultados, e consegue processar 1000 imagens por segundo, ou seja, demora 1 milissegundo por cada imagem. Assim, a implementação deste método disponibilizada na biblioteca Dlib será utilizada na solução a desenvolver.

3.6 Conclusão

Neste capítulo foram avaliados alguns dos métodos enunciados no estado da arte, e foi escolhido o método base para a solução a desenvolver. Foram também apresentadas as limitações principais desse método, juntamente com algumas soluções para essas limitações.

4 *Design* da Solução

Este capítulo apresenta o *design* da solução desenvolvida e as tecnologias usadas.

4.1 Arquitetura

A solução desenvolvida será uma aplicação de *desktop* e terá uma arquitetura simples de duas camadas, a da interface gráfica e a da lógica de negócio (Figura 9). A camada da interface gráfica trata de mostrar ao utilizador o vídeo original, a dobragem, e o resultado final, e processa as ações do utilizador na interface gráfica. A camada da lógica de negócio contém todas as partes relativas à captura e reprodução dos movimentos faciais, como as entidades (e.g. atores) e algoritmos (e.g. captura de movimentos e síntese do interior da boca).

Uma arquitetura de Cliente-Servidor, embora seja uma alternativa possível, iria introduzir uma demora na execução do algoritmo, já que teria de enviar e receber as imagens, e implicaria um aumento dos recursos computacionais do servidor de forma a conseguir processar os pedidos das várias aplicações. Uma aplicação de *desktop* consegue aceder com facilidade às capacidades completas do computador.

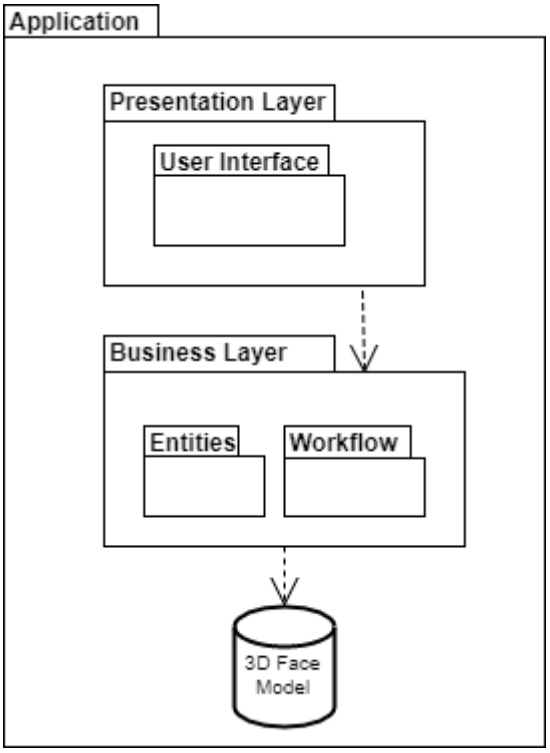

Figura 9: Arquitetura da aplicação, com duas camadas

A solução poderia funcionar como uma extensão (*add-on*) às aplicações de dobragem, mas seria difícil suportar o grande número de aplicações existentes. Ao ser uma aplicação normal, funciona em paralelo com as gravações das dobragens, e o vídeo resultante pode ser colocado no filme dobrado utilizando *software* especializado.

O modelo 3D é lido diretamente para as estruturas de dados respetivas, durante a inicialização da aplicação. Assim, não é preciso uma camada de acesso aos dados, como aconteceria se estivesse a ser usado um sistema de gestão de bases de dados.

A Figura 10 contém um diagrama das classes da solução desenvolvida. A ligação entre a camada da interface gráfica e a camada da lógica de negócio é efectuada num só ponto, *ApplicationController*. Esta classe trata de processar as imagens, passando-as pela deteção da face (*DlibFaceDetector*) e dos pontos faciais (*DlibLandmarkDetector*). A criação do modelo 3D da face é feita por uma biblioteca externa. A transferência dos movimentos para o modelo do ator original é feita pela classe *DeformationTransferManager*. A síntese da boca é efetuada pela classe *MouthSynthesizer* (Figura 11). Finalmente, a criação da imagem final é efetuada pela classe *ApplicationController* no final de cada iteração.

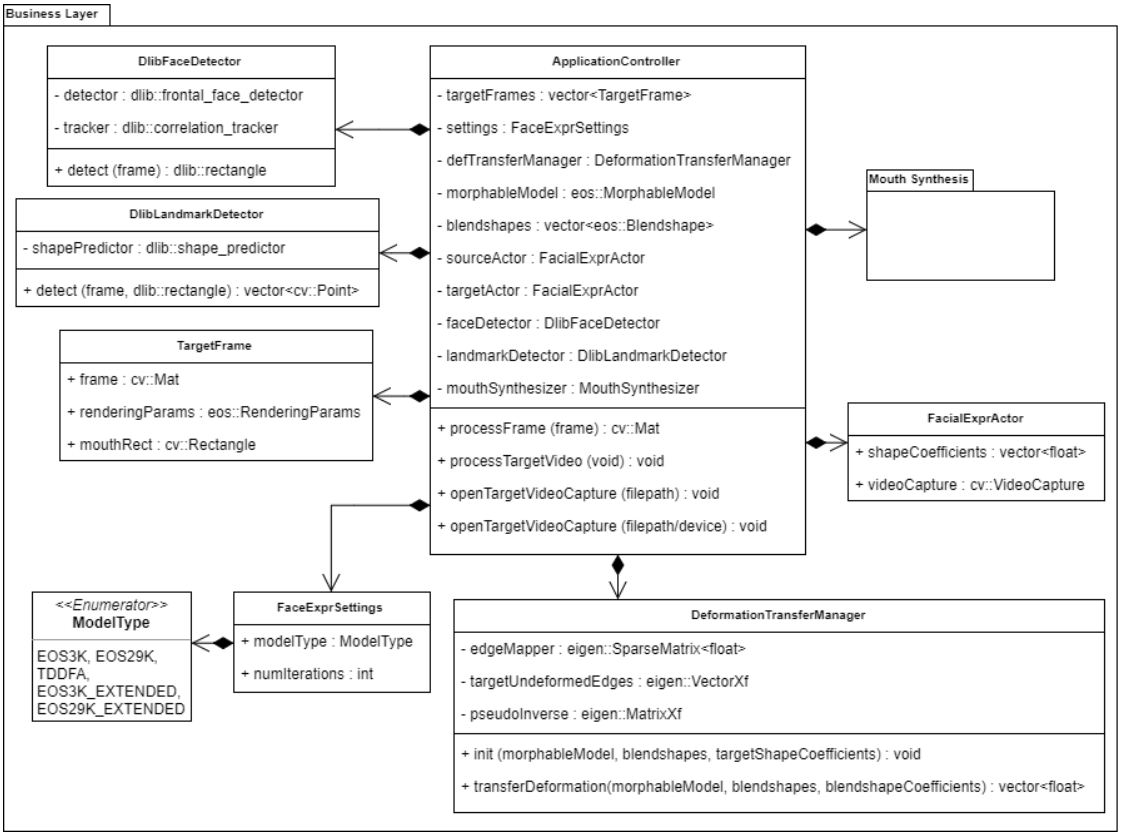

Figura 10: Diagrama de classes da solução

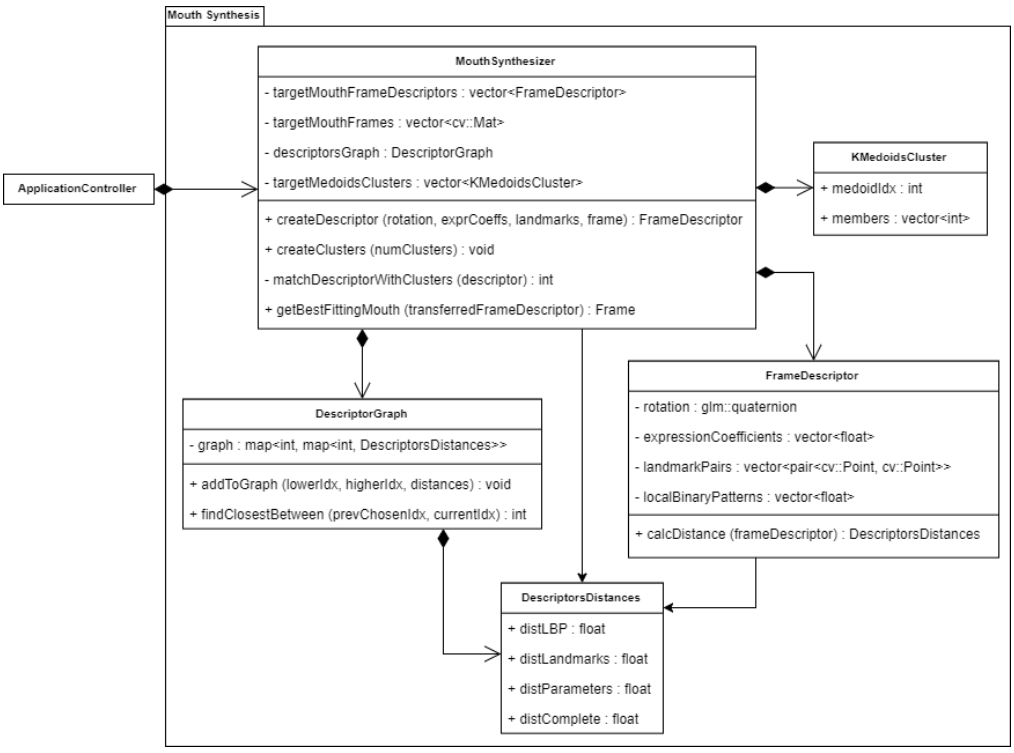

Figura 11: Diagrama de classes da síntese da boca

4.2 Linguagens e Tecnologias

A linguagem de programação utilizada foi C++, pela variedade de bibliotecas úteis disponíveis, e porque é uma linguagem frequentemente utilizada em aplicações de execução em tempo real. Alternativas incluem C# e Java, entre outras.

Para o pré-processamento das imagens, aquisição das imagens, e composição da imagem final, foi utilizada a biblioteca open-source OpenCV[6], que é provavelmente a maior e mais utilizada biblioteca na área, com mais de 2500 algoritmos optimizados. Opções alternativas incluem BoofCV[7] (Java), EmguCV[8] (*wrapper* de OpenCV para C#), e AForge.NET[9] (C#). OpenCV também inclui estruturas de dados úteis, como matrizes (cv::Mat), vectores, pontos (cv::Point), entre outras.

Foi utilizada a biblioteca Dlib dada a qualidade e velocidade dos algoritmos de deteção de caras e de detecção dos pontos faciais e a fácil integração com OpenCV.

Foram utilizadas as bibliotecas GLM[10] e Eigen[11], pelas suas estruturas de dados e algoritmos relacionados com álgebra linear.

4.3 Interface Gráfica

A interface gráfica é uma interface simples, como se pode ver na Figura 12. Permite a escolha e visualização do vídeo não dobrado, e a visualização simultânea do ator e do resultado da dobragem. Após a dobragem, também é possível visualizar o resultado, para confirmar que está correcto. Finalmente, é possível exportar o resultado para um ficheiro.

Para a criação da interface gráfica foi utilizada a *framework* multiplataforma Qt, que funciona bem com arquitetura escolhida, já que deixa que exista uma separação entre a camada da interface gráfica e a lógica de negócio. Alternativas incluem wxWidgets (também para C++), WindowsForms (C#), e Swing (Java).

[6] http://opencv.org/
[7] http://boofcv.org/
[8] http://www.emgu.com/
[9] http://www.aforgenet.com/
[10] https://glm.g-truc.net/
[11] http://eigen.tuxfamily.org/

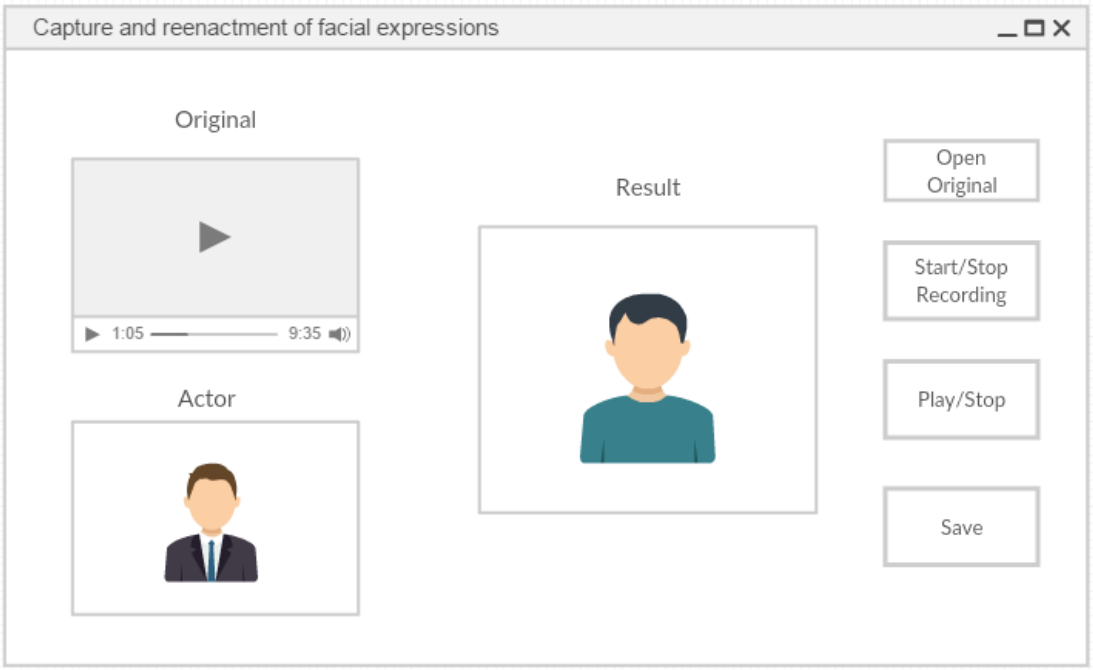

Figura 12: *Mock-up* da interface gráfica

4.4 Conclusão

Neste capítulo foi apresentado o *design* da solução desenvolvida, com um diagrama de arquitetura e diagramas de classe. Foram também apresentadas as tecnologias usadas no desenvolvimento.

5 Construção da Solução

Este capítulo descreve a construção da solução, seguindo o *design* apresentado no capítulo 4. São detalhados os algoritmos implementados e é descrita a forma como estes integram a solução completa.

5.1 Modelo 3D

O algoritmo de captura e transferência de movimentos faciais tem como estrutura base um modelo 3D de faces, com blendshapes de expressões/movimentos faciais. Este modelo 3D genérico é obtido a partir da geometria facial de várias pessoas. Ao conjunto de geometrias faciais é aplicada a análise de componentes principais (PCA, *Principal Component Analysis*), de forma a poder reduzir a dimensionalidade do modelo (por exemplo, em vez de utilizar as geometrias faciais de 200 pessoas, podem ser utilizados apenas 60 componentes principais).

O modelo utilizado é o modelo de geometria facial BFM (*Basel Face Model*) de Paysan et al. (2009), modificado por Tran et al. (2016), com 29 *blendshapes* obtidas a partir do modelo de Cao, Weng, et al. (2014). Com 46990 vértices, este modelo 3D revelou-se demasiado grande e resultava numa captura e transferência demasiado lenta. Por isso, utilizando um programa externo[12], as *blendshapes* foram transferidas para um modelo 3D de resolução mais baixa, disponibilizado publicamente com a biblioteca eos[13] (Huber et al. 2016). Entre as várias resoluções disponíveis, foram utilizadas as de 3448 vértices e de 29587 vértices.

[12] https://github.com/Golevka/deformation-transfer
[13] https://github.com/patrikhuber/eos

Figura 13: Exemplos dos modelos 3D utilizados. Em cima, a geometria facial base. Em baixo, a geometria facial com uma só *blendshape*. Da esquerda para a direita: o modelo usado inicialmente, o modelo de 3448 vértices, e o modelo de 29587 vértices

Num modelo deste tipo, uma malha triangular com certos movimentos faciais pode ser obtida com a seguinte fórmula:

$$M(\alpha, \beta) = \mu_{shape} + PC_{shape} \cdot \alpha + PC_{exp} \cdot \beta \quad (1)$$

, onde $\mu_{shape} \in \mathbb{R}^{3n}$ é a geometria facial média, $PC_{shape} \in \mathbb{R}^{3n \times 63}$ é a matriz com os componentes principais da geometria facial, $\alpha \in \mathbb{R}^{63}$ é um vetor com os coeficientes da geometria facial, $PC_{exp} \in \mathbb{R}^{3n \times 29}$ é uma matrix com as *blendshapes* dos movimentos faciais, e $\beta \in \mathbb{R}^{29}$ é um vetor com os coeficientes das *blendshapes*, sendo n o número de vértices do modelo. Alterando os coeficientes da geometria facial e das *blendshapes*, é possível gerar várias faces (Figura 13).

5.2 Algoritmo

O algoritmo desenvolvido pode ser dividido em duas partes principais, uma para cada ator, com alguns passos em comum. Inicialmente, é executada a parte do ator alvo (o ator original, que será alterado de acordo com a dobragem) para se obter a informação necessária, como a geometria facial. Após se processar o vídeo desse ator, é executada a parte relativa ao ator da dobragem, onde se captura os movimentos a transferir e se efetua essa transferência para o ator alvo. As secções seguintes descrevem o algoritmo mais detalhadamente.

5.2.1 Passos comuns

O algoritmo (Figura 14), em ambas as partes, começa por procurar uma face na imagem que recebe (por exemplo, de um vídeo ou de uma câmara) e após a encontrar, procede à deteção de pontos faciais nessa face. Para simplificar o algoritmo, é usada a maior face que o detetor encontra.

Utilizando os pontos faciais detetados, efetua-se a otimização do modelo 3D genérico para obter a geometria facial, a pose (rotação da cabeça), e movimentos faciais, através do algoritmo da biblioteca eos. Esta optimização é diferente consoante o algoritmo esteja na fase de inicialização ou na fase após a inicialização. A fase de inicialização ocorre nas primeiras N imagens, onde N é igual a seis, baseado no número utilizado por outros métodos, embora possa ser utilizado um número diferente. Nesta fase, é feita a otimização completa (geometria, pose, e movimentos faciais), e na fase seguinte é utilizada uma geometria fixa, e só é feita a otimização da pose e dos movimentos faciais. Os coeficientes da geometria fixa são a média dos coeficientes obtidos nas N imagens. Esta diferença entre as duas fases proporciona uma execução mais rápida na segunda fase, e a existência de uma geometria fixa pode ser aproveitada em várias partes do algoritmo, nomeadamente na transferência dos movimentos.

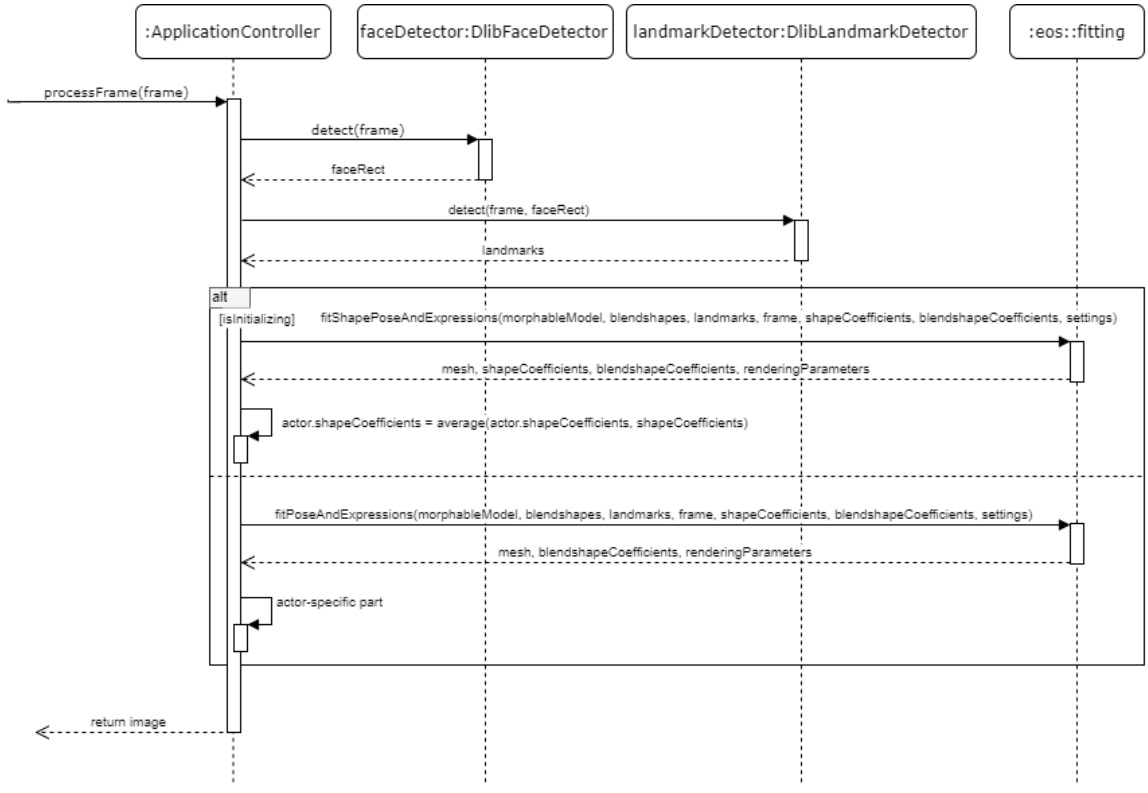

Figura 14: Diagrama de sequência dos passos comuns aos dois atores

5.2.2 Captura da geometria e movimentos faciais

Para obter a geometria facial, a rotação da cabeça e os movimentos faciais de um ator numa imagem, foi utilizado o algoritmo da biblioteca eos. Na fase de inicialização, é executado o algoritmo completo: otimização da rotação da cara utilizando os pontos faciais, incluindo os do contorno do rosto, otimização da geometria facial a esses pontos utilizando a rotação obtida, e otimização dos movimentos faciais utilizando a rotação e geometria obtidas. Esta sequência de passos é repetida várias vezes (cinco iterações por exemplo), para se obter as estimações mais correctas. O algoritmo é iniciado com os coeficientes todos a zero, ou com os obtidos na imagem anterior, e cada iteração utiliza os coeficientes e rotação obtidos da iteração anterior.

Após a fase de inicialização, é utilizada uma versão modificada deste algoritmo, que só obtém a rotação e os coeficientes dos movimentos faciais, utilizando uma geometria fixa.

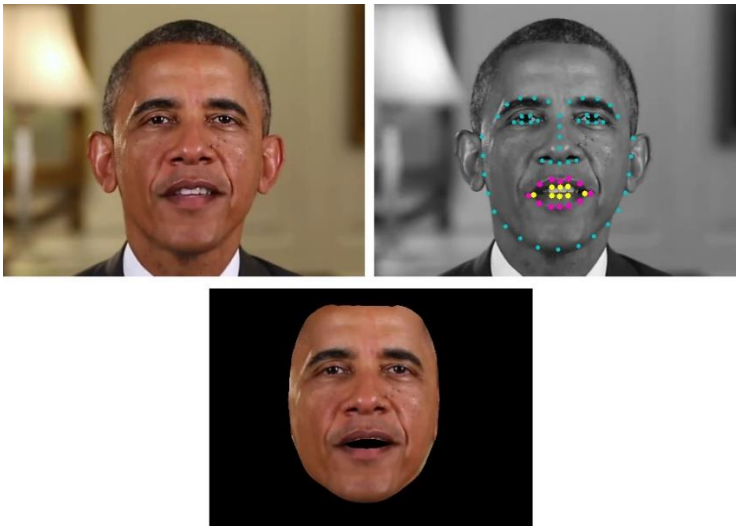

Figura 15: A imagem inicial, os pontos faciais detetados, e a geometria facial obtida (com a textura sobreposta)

5.2.3 Ator alvo

Na fase a seguir à inicialização, cada imagem é processada para obter a respectiva informação sobre o interior da boca, e são guardadas numa lista a imagem, a pose obtida, e o resto da informação necessária para utilizar cada imagem na fase de composição final da imagem alterada (Figura 16).

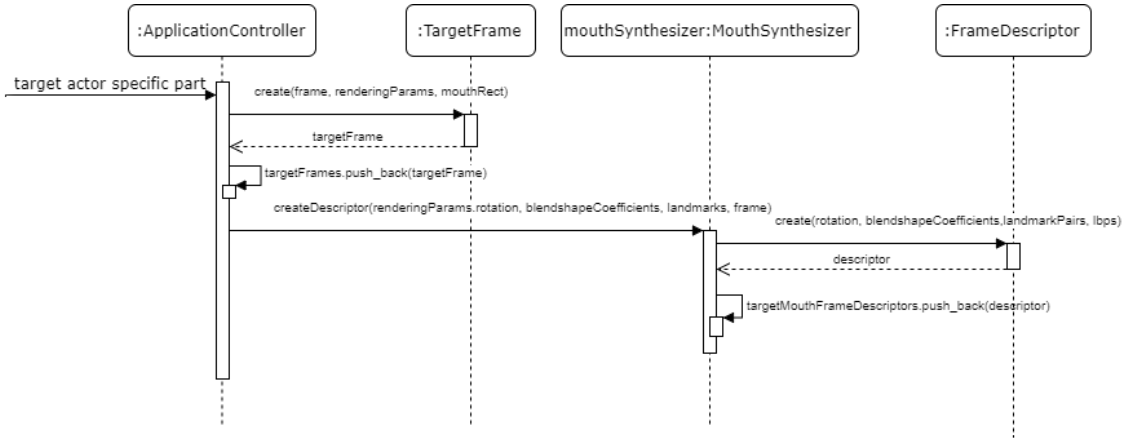

Figura 16: Diagrama de sequência dos passos específicos do ator alvo

5.2.4 Ator dobrador

Logo a seguir à fase de inicialização, é feita a inicialização do algoritmo de transferência dos movimentos faciais. Na fase seguinte, os movimentos são transferidos para a geometria facial do ator alvo, obtendo-se os coeficientes dos movimentos faciais transferidos. A seguir é obtido o interior da boca que mais se adequa a esses coeficientes, e por fim é efetuada a composição de três imagens: o interior da boca, a renderização da face do ator alvo com os movimentos faciais transferidos, e a imagem do vídeo original (Figura 17).

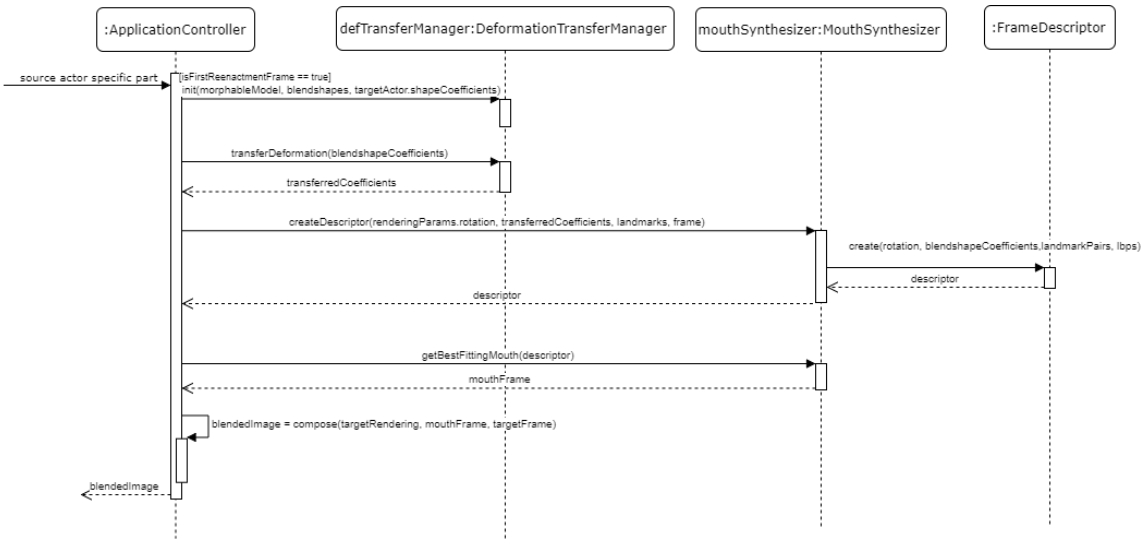

Figura 17: Diagrama de sequência dos passos específicos do ator dobrador

5.2.5 Transferência de movimentos faciais

A transferência de movimentos faciais tem como objetivo alterar a malha triangular base (geometria facial) do ator alvo para que esta reproduza as alterações que ocorreram entre a malha base e a malha com movimentos faciais. Aqui, a malha base é a da geometria facial numa posição neutra, ou seja, sem qualquer movimento facial. Na literatura, a malha base/neutra é considerada a malha não deformada, e a malha com movimentos faciais é a malha deformada, onde por deformação entende-se a alteração das posições de vértices da malha não deformada.

O algoritmo de transferência de expressões faciais implementado (Thies et al. 2016) foi baseado no método de transferência de deformações entre duas malhas triangulares, de Sumner & Popović (2004). Esta transferência de deformações baseia-se na utilização de gradientes de deformação, que quando são aplicados à malha não deformada, resultam numa malha deformada. Utilizando os vetores representativos de duas arestas do triângulo j cujos vértices têm os índices $\{i_1, i_2, i_3\}$:

$$\begin{aligned} V_j &= [v_{i_2} - v_{i_1} \quad v_{i_3} - v_{i_1}], \\ \tilde{\mathrm{V}}_j &= [\tilde{\mathrm{v}}_{i_2} - \tilde{\mathrm{v}}_{i_1} \quad \tilde{\mathrm{v}}_{i_3} - \tilde{\mathrm{v}}_{i_1}] \end{aligned} \tag{2}$$

, onde os vértices v e $\tilde{\mathrm{v}}$ pertencem, respetivamente, à malha não deformada e à malha deformada. Assim, o gradiente de deformação J_j é a matrix 3×3 que deforma o triângulo j:

$$J_j \cdot V_j = \tilde{\mathrm{V}}_j \tag{3}$$

A teoria por detrás dos gradientes de deformação, e mais informação sobre o método de transferência, pode ser encontrada na dissertação de Sumner (2005), de onde estas equações foram retiradas.

Tendo a malha não deformada do ator alvo e a do ator dobrador, e a malha deformada do ator dobrador, pretende-se obter a malha deformada do ator alvo. Idealmente, a deformação será a mesma que a do ator dobrador, por isso são aplicados os gradientes de deformação à malha do ator alvo, obtendo assim a malha deformada ideal. O objetivo será obter a malha deformada do ator alvo que mais se aproxime desta malha ideal:

$$\min \sum_{j=1}^{|F|} \left\| J_j \cdot V_j - \tilde{\mathrm{V}}_j \right\|_F^2 \tag{4}$$

, onde $|F|$ é o número de triângulos da malha, J é a matriz com os gradientes de deformação de todos os triângulos, V e $\tilde{\mathrm{V}}$ são matrizes com os vetores das arestas de todos os triângulos, da malha não deformada e da malha deformada, respetivamente, e $\|\cdot\|_F$ é a norma matricial. J é calculado em cada nova imagem, de acordo com a equação 3, após a captura dos movimentos faciais do ator dobrador. Este problema de minimização é resolvido através do método dos mínimos quadrados lineares[14].

Seguindo a equação 1, a malha deformada depende de coeficientes desconhecidos (que queremos obter aqui), das blendshapes de movimentos faciais, e da malha não deformada. Assim, a matriz de arestas da malha deformada será:

$$\tilde{\mathrm{V}} = EdgeMap \cdot \left(PC_{exp} \cdot \beta + NeutralMesh \right) \tag{5}$$

[14] https://en.wikipedia.org/wiki/Linear_least_squares_(mathematics)

, onde $NeutralMesh$ é a malha sem movimentos faciais (só a geometria base), e $EdgeMap \in \mathbb{R}^{6|F|\times 3n}$ é uma matriz, obtida no início da aplicação, que faz o mapeamento entre uma malha e a matriz de arestas V, segundo a equação 2, e que tem o seguinte formato:

$$EdgeMap = \begin{bmatrix} 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \dots & \dots & \dots & \dots & -1 & \dots & \dots & \dots \\ \dots & \dots & 1 & \dots & \dots & \dots & \dots & -1 & \dots & \dots \\ \dots & \dots & \dots & 1 & \dots & \dots & \dots & \dots & -1 & \dots \\ -1 & \dots & \dots & \dots & \dots & \dots & \dots & 1 & \dots & \dots \\ \dots & -1 & \dots & \dots & \dots & \dots & \dots & \dots & 1 & \dots \\ \dots & \dots & -1 & \dots & \dots & \dots & \dots & \dots & \dots & 1 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \end{bmatrix} \quad (6)$$

, ou seja, uma matriz esparsa, com apenas dois elementos (1 e -1) por cada linha. Assim, cada três linhas de V contêm as coordenadas de um vetor de uma aresta.

Modificando a equação 4 de acordo com a equação 5, 5555555obtém-se:

$$\min \left\| J \cdot V - EdgeMap \cdot \left(PC_{exp} \cdot \beta + NeutralMesh \right) \right\|_2^2 \quad (7)$$

Esta equação pode ser simplificada para obter uma nova equação a minimizar:

$$\min \left\| (J \cdot V - EdgeMap \cdot NeutralMesh) - \left(EdgeMap \cdot PC_{exp} \right) \cdot \beta \right\|_2^2 \quad (8)$$

, que segue o formato $\|b - Ax\|_2^2$. O resultado de uma minimização deste tipo, com mais linhas (seis por cada triângulo) do que colunas (número de *blendshapes*), pode ser aproximado com as equações normais:

$$Ax = b \Leftrightarrow A^T Ax = A^T b \Leftrightarrow x = (A^T A)^{-1} A^T b \quad (9)$$

$(A^T A)^{-1} A^T$ é a pseudo-inversa[15] A^+ , resultando em:

$$x = A^+ \, b \quad (10)$$

A pseudo-inversa é calculada através da decomposição em valores singulares (SVD, *Singular Value Decomposition*), após a fase de inicialização do ator dobrador. Para transferir os movimentos faciais, b é calculado com a malha deformada do ator dobrador, e é resolvida a equação 10, obtendo os coeficientes dos movimentos faciais na malha do ator alvo.

5.2.6 Síntese do interior da boca

Para obter o interior da boca que mais se adequa aos movimentos faciais transferidos, foi implementado o algoritmo de Thies et al. (2016). Após a fase de inicialização do ator alvo, é guardada a informação do interior da boca de cada imagem. Esta informação inclui: rotação R da cara, coeficientes β dos movimentos faciais, pontos faciais F, e os padrões binários locais (LBP, *Local Binary Patterns*) L da boca. Obtém-se assim um descritor para cada boca, $K = \{R, \beta, F, L\}$.

[15] https://en.wikipedia.org/wiki/Moore%E2%80%93Penrose_inverse

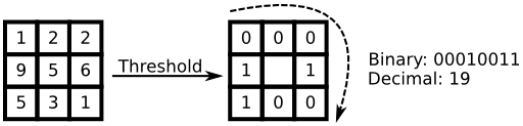

Figura 18: Exemplo de um padrão binário local.
Fonte:(Wagner 2011)

O padrão binário local[16] é um descritor muito utilizado para classificação, e o tipo mais básico consiste em comparar os pixéis à volta de um pixel, no sentido dos ponteiros, e caso o valor de um pixel seja menor do que o do pixel vizinho, escreve-se um "0", caso contrário escreve-se um "1". Obtém-se assim um padrão, como "00010011" (Figura 18). Criando um histograma de todos os padrões, obtém-se um histograma que descreve a imagem em termos dos padrões binários. Dividindo a imagem numa grelha, e concatenando os histogramas de cada grelha, consegue-se também incluir informação espacial. Seguindo Garrido et al. (2013), a imagem da boca é dividida numa grelha de três linhas e cinco colunas (Figura 19), e para cada quadrado são obtidos os histogramas de padrões binários com vizinhança de oito e de quatro.

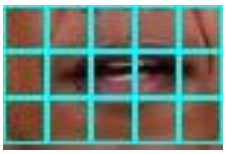

Figura 19: Exemplo da grelha utilizada para os padrões binários locais

A distância (ou diferença) entre dois descritores K é definida da seguinte forma (equações adaptadas de Thies et al. (2016)):

$$D(K_1, K_2) = D_p(K_1, K_2) + D_M(K_1, K_2) + D_a(K_1, K_2) \quad (11)$$

, onde D_p é a distância das rotações e dos coeficientes:

$$D_p(K_1, K_2) = \|\beta_1 - \beta_2\|_2^2 + \|R_1 - R_2\|_F^2 \quad (12)$$

, e D_m é a distância relativa aos pontos faciais da boca:

$$D_m(K_1, K_2) = \sum_{(i,j)\in\Omega} \left(\left\| F_{1_i} - F_{1_j} \right\|_2 - \left\| F_{2_i} - F_{2_j} \right\|_2 \right)^2 \quad (13)$$

, onde Ω é um conjunto de pares de pontos faciais, ou seja, pontos do lábio superior com os seus equivalentes no lábio inferior, e pontos dos cantos da boca. Finalmente, D_a é a distância qui-quadrada dos histogramas de padrões binários locais L.

Para obter o interior de boca mais adequado aos movimentos faciais transferidos, é criado um descritor K com os coeficientes obtidos na transferência dos movimentos, e com os pontos faciais do ator dobrador. Como seria muito demorado comparar este descritor com todos os descritores do ator alvo, e como muitos dos descritores serão similares entre si (por exemplo, o descritor de uma imagem será similar ao da imagem seguinte), é necessário cortar no número de comparações a fazer. Para isso, é utilizado uma versão modificada do método de agrupamento (*clustering*) *k-means*, conhecida por *k-medoids*. Enquanto que no método *k-means* o agrupamento em k grupos é feito em relação a médias, no método *k-medoids* os

[16] https://bytefish.de/blog/local_binary_patterns/

descritores são agrupados em relação ao descritor mais perto. Após a fase de inicialização do ator alvo, os descritores são agrupados em 10 grupos, em que o representante (centro) de cada grupo é o descritor com a distância mínima a todos os descritores do seu grupo (Figura 20).

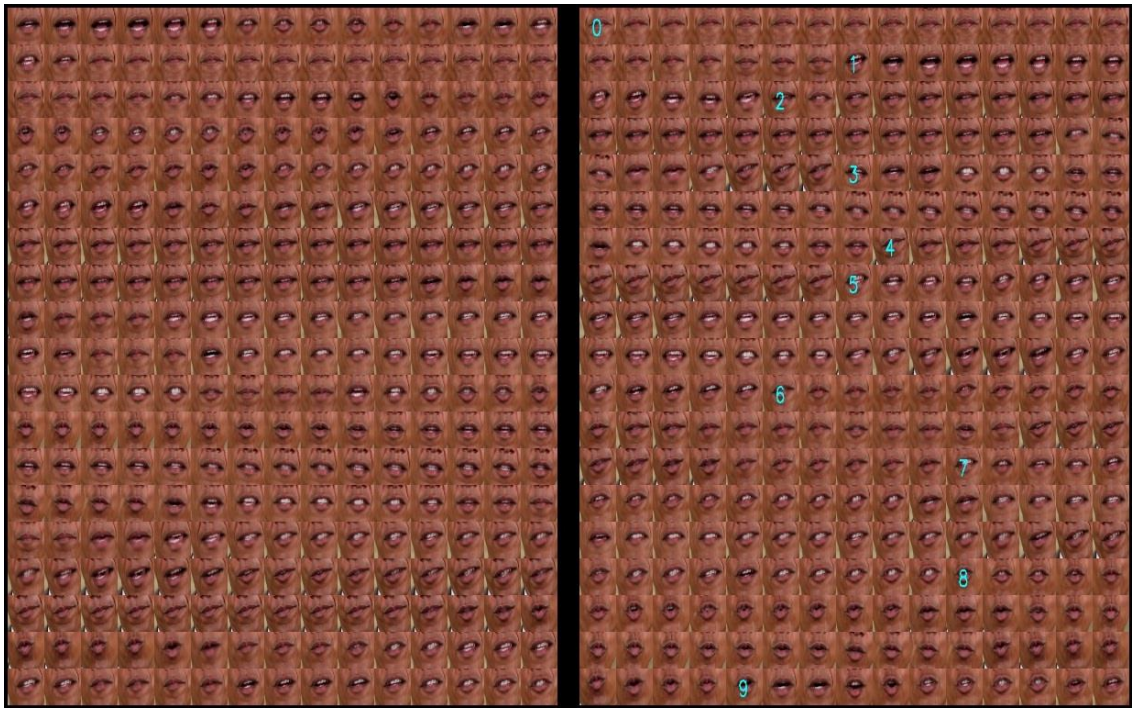

Figura 20: À esquerda, imagens da boca obtidas em 300 *frames* de um vídeo. À direita, as mesmas imagens ordenadas pelo grupo em que ficaram após o método *k-medoids*

Para obter o descritor do interior da boca mais adequado, o descritor K é comparado aos descritores representantes dos grupos, e é escolhido o representante que esteja mais perto. Utilizando a métrica de distância entre dois descritores, é criado um grafo completo de todos os descritores, onde o valor de cada aresta é a distância entre os descritores que liga. Após obter o descritor representante mais perto de K, obtém-se o descritor que minimiza a soma das arestas do grafo até ao representante escolhido e até ao descritor escolhido na imagem anterior. Este descritor é o que representa o melhor interior de boca para os movimentos transferidos na imagem corrente.

5.2.7 Composição final

Na parte final existem três imagens separadas: o interior da boca correspondente ao descritor escolhido, a imagem sintetizada da face do ator alvo com os movimentos faciais transferidos, e a imagem do vídeo original. A imagem do interior da boca é deformada através de uma transformação afim para que esteja ajustada à posição da boca na malha deformada do ator alvo. Essa imagem é combinada com a imagem do ator alvo com os movimentos transferidos, e a imagem resultante é combinada com a imagem do vídeo original (Figura 21). Em ambas as

combinações é efetuado um alisamento da imagem nas partes em que as imagens separadas se encontram, de forma a eliminar artefactos nessas partes.

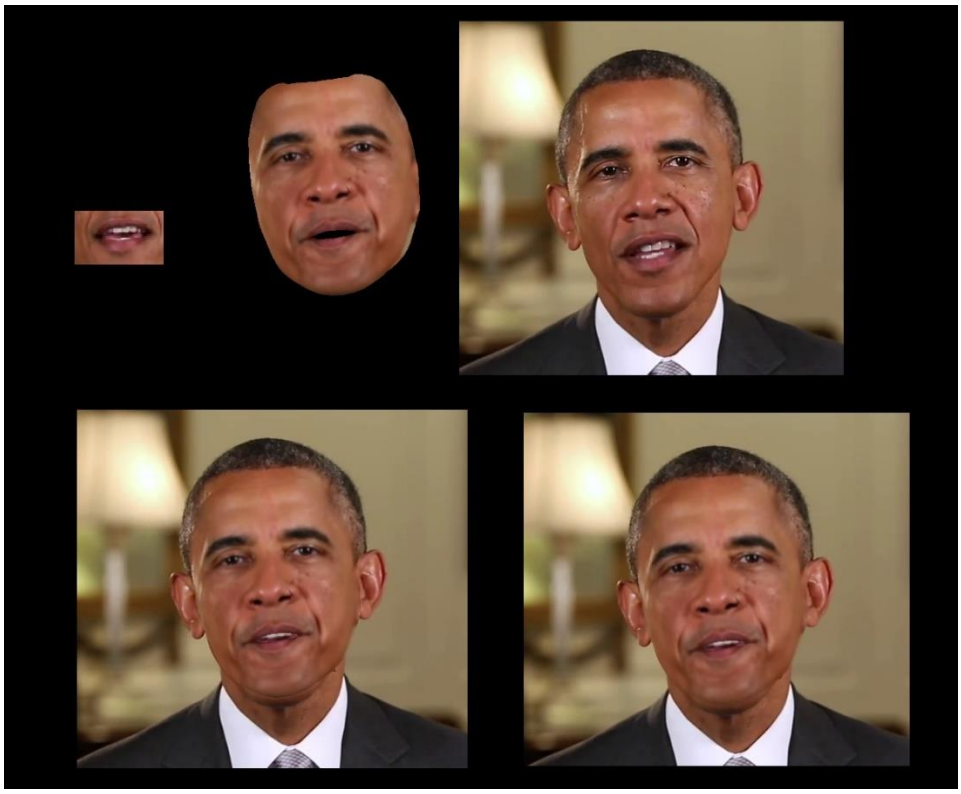

Figura 21: Exemplo da composição das três imagens. Em baixo: à esquerda, a imagem composta, sem alisamento, e à direita a mesma imagem, com alisamento.

5.3 Conclusão

Neste capítulo foram descritos detalhadamente os algoritmos implementados, como o de transferência de movimentos faciais, e o método de composição da imagem final após a captura e transferência de movimentos faciais do ator da dobragem para o ator alvo.

6 Avaliação da Solução

Este capítulo descreve a forma como a solução desenvolvida foi avaliada e comparada com as soluções existentes. A solução desenvolvida foi avaliada de duas formas: pela qualidade (exatidão) do resultado, e pela satisfação dos espetadores em relação às dobragens tradicionais.

6.1 Avaliação da Exatidão

A avaliação da exatidão das imagens sintetizadas, ou seja, do resultado final do algoritmo, tem como objetivo avaliar se a solução desenvolvida é melhor do que as soluções existentes, em termos de realismo do resultado final. Uma imagem com artefactos na face, como lábios deformados, será menos exacta que uma sem esses artefactos, logo será menos realista. Um dos objetivos principais destas soluções é aumentar o realismo das dobragens, por isso esta é uma avaliação necessária.

Como a solução descrita em (Thies et al. 2016) pode ser considerada a de maior qualidade neste momento (de acordo com os resultados demonstrados pelos autores), a hipótese a testar é "A solução desenvolvida é melhor que a solução descrita em (Thies et al. 2016)".

Esta exatidão é avaliada comparando a imagem sintetizada a uma imagem real. Para se obter uma imagem sintetizada referente os mesmos movimentos que uma imagem real, ambos os atores (original e de dobragem) têm de ser a mesma pessoa, e a cena (lugar, iluminação, roupa, etc.) também tem de ser igual. Uma forma fácil de conseguir isto consiste em utilizar o mesmo vídeo como ator original e como ator de dobragem.

Para medir a diferença entre as duas imagens pode-se utilizar o fluxo óptico, que obtém o deslocamento de cada pixel de uma imagem para a outra (Figura 22). A média destes deslocamentos pode ser considerada a medida de exatidão da solução, e quanto menor for a média, maior é a exatidão. Idealmente não haverá deslocamento dos pixels, o que significa que a solução sintetizou uma imagem completamente igual à original.

Figura 22: Da esquerda para a direita: original, resultado, e deslocamento de acordo com o fluxo óptico. Fonte: material suplementar de (Thies et al. 2016)

Os autores de (Thies et al. 2016) publicaram apenas a avaliação de um vídeo, pelo que não é possível obter uma comparação extensa entre os resultados deles e os resultados obtidos pela solução desenvolvida. A Tabela 1 apresenta os resultados obtidos por Thies et al. (2016) e pela solução, relativos ao mesmo vídeo.

Tabela 1 – Resultados de Thies et al. (2016) e da solução desenvolvida

	Média (em píxeis)	Desvio padrão (em píxeis)
Thies et al (2016)	0.33	0.157
Solução desenvolvida	1.26	7.83

Os resultados apresentados demonstram que a solução desenvolvida não conseguiu obter resultados melhores que os obtidos por Thies et al. (2016), pelo menos no vídeo analisado.

6.2 Avaliação da Satisfação dos Espetadores

A segunda grandeza a avaliar é a satisfação dos espetadores quanto à dobragem feita utilizando a solução desenvolvida. É importante saber qual a opinião dos espetadores sobre a dobragem, em relação às dobragens tradicionais, nomeadamente saber se a solução conseguiu reduzir a sensação de anormalidade e se eles preferem as novas dobragens ou as tradicionais.

A hipótese a testar é "Os espetadores estão mais satisfeitos com as dobragens feitas com a solução desenvolvida do que com as dobragens tradicionais". Aqui o objetivo é saber como é que as novas dobragens se comparam às tradicionais.

Para avaliar esta grandeza, foi criado um questionário de satisfação, que foi colocado a espetadores após visualizarem dois pares de dois vídeos, em que cada par consistia num vídeo com uma dobragem tradicional e noutro com a nova dobragem. Os vídeos foram originalmente gravados em alemão, e posteriormente dobrados em inglês por um estúdio profissional de dobragem.

Figura 23: Imagens dos vídeos utilizados no questionário. Da esquerda para a direita: Vídeo 1 (dobragem tradicional), Vídeo 2 (solução), Vídeo 3 (dobragem tradicional) e Vídeo 4 (solução)

As questões colocadas estavam relacionadas com a qualidade da dobragem, e a sensação de anormalidade, entre outras. As respostas estavam numa escala de Likert (e.g. desde "não concordo totalmente" a "concordo totalmente"), e foram associados valores numéricos a cada resposta. Para cada respondente foram somados os valores das respostas, obtendo um valor equivalente a uma pontuação (Figura 24). Com cinco perguntas em cada video, cada vídeo tem uma pontuação mínima de cinco valores, e uma pontuação máxima de 25 valores.

Foram recolhidas respostas de 23 pessoas, todas de nacionalidade portuguesa, com idades entre os 15 e os 38 anos (Figura 25). Com unanimidade, todos responderam que preferiam legendagens, e não dobragens, confirmando o que foi mencionado relativamente às preferências dos portugueses.

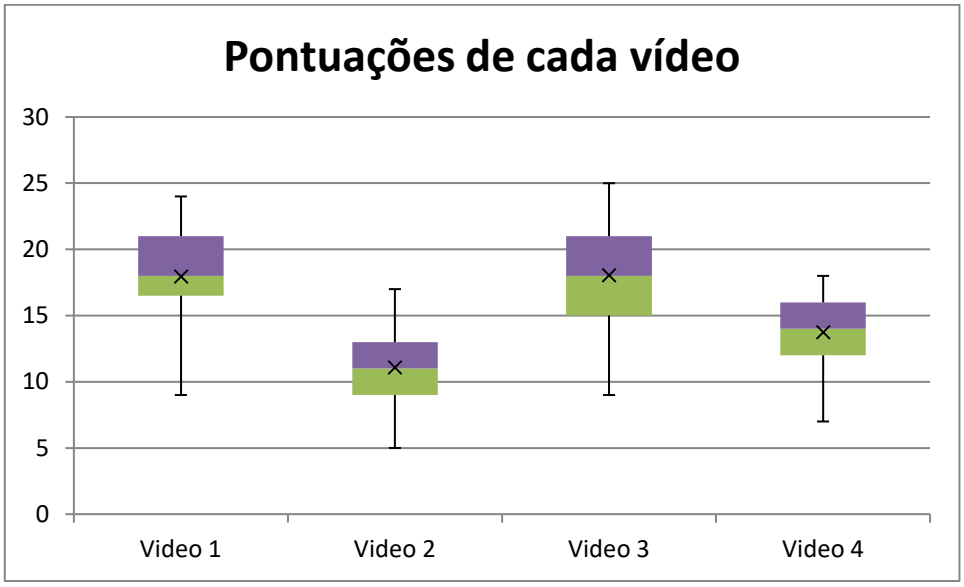

Figura 24: Gráfico de caixa das pontuações de cada vídeo

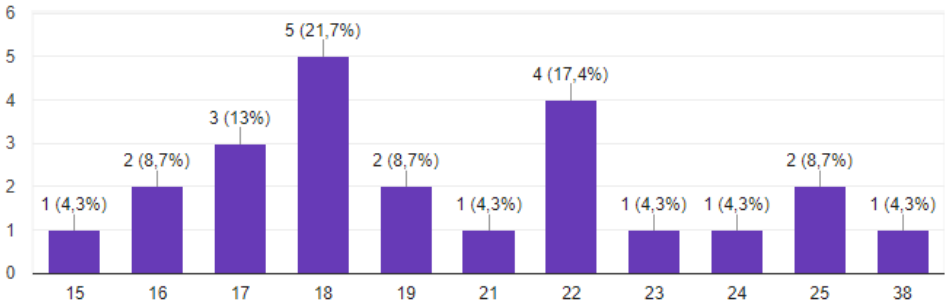

Figura 25: Gráfico das idades dos respondentes

Foi efetuado o teste de Shapiro-Wilk, para verificar se os dados recolhidos pertenciam a uma distribuição normal, e a hipótese nula de os dados não pertencerem a uma distribuição normal foi rejeitada em ambos os pares. Como para cada par de vídeos existem dois questionários a comparar (um sobre a dobragem tradicional e o outro sobre a dobragem feita pela solução desenvolvida), feitos aos mesmos respondentes, e pertencendo a uma distribuição normal, foi utilizado o teste t de Student (*paired samples*). Seguindo as hipóteses:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 > \mu_2$$

, queremos rejeitar a hipótese nula das médias das pontuações serem iguais entre os questionários. Como foram obtidos *p-values* menores que 0.05, a hipótese nula é rejeitada em ambos os pares, e existe uma diferença significativa entre as pontuações (Tabela 2 e

Tabela 3).

Tabela 2 – *p-values* dos Video 1 e 2, obtidos com o *add-in* RealStatistics[17], para Excel

	p-value	t-crit	lower	upper	sig
One Tail	9,9E-08	1,717144			yes
Two Tail	2E-07	2,073873	4,951981	8,787149828	Yes

Tabela 3 – *p-values* dos Video 3 e 4

	p-value	t-crit	lower	upper	sig
One Tail	0,000102	1,717144			yes
Two Tail	0,000204	2,073873	2,295282	6,313413	yes

As médias de pontuações das dobragens tradicionais (17.96 e 18.04) foram superiores às das dobragens criadas pela solução desenvolvida (11.09 e 13.74). Quando questionados sobre que vídeos preferiam, maior parte escolheu a dobragem tradicional, embora alguns tenham escolhido a dobragem criada pela solução. Assim, podemos concluir que os espetadores não estão mais satisfeitos com as dobragens criadas pela solução, o que se deve a vários fatores, nomeadamente os que já foram mencionados no capítulo 3. Convém mencionar que as dobragens tradicionais não obtiveram pontuações perfeitas, o que reforça a necessidade de melhorarmos as dobragens.

6.3 Conclusão

Neste capítulo foi descrita a forma como a solução desenvolvida foi avaliada, e como foram analisados os resultados de comparações com o método base e de questionários de satisfação de espetadores. Analisando os resultados obtidos em ambas as formas de avaliação, podemos concluir que a solução desenvolvida não obteve nem resultados melhores que o estado da arte nem resultados melhores que as dobragens tradicionais.

[17] http://www.real-statistics.com/

7 Conclusão

Esta dissertação tinha como objetivo o desenvolvimento uma solução de captura e reprodução de movimentos faciais, para capturar os movimentos faciais de um ator de dobragem e fazer com que um ator previamente gravado reproduzisse esses movimentos. Com esta solução pretendia-se eliminar alguns dos problemas existentes com dobragens tradicionais, de forma a tornar os resultados mais realistas, diminuir a discrepância entre as partes visuais e auditivas (que podem resultar em falhas na compreensão) e melhorar a experiência de visualização por parte dos espetadores.

Para atingir este objetivo, foi avaliado o estado da arte de métodos de captura e reprodução de movimentos faciais, e foi desenvolvida uma solução baseada nesse estado da arte. A solução desenvolvida captura a geometria facial e os movimentos faciais de dois atores, um ator alvo e um ator dobrador, e efetua a transferência dos movimentos do ator dobrador para a geometria facial do ator alvo. Após a transferência, é obtido o interior da boca que mais se adequa aos movimentos transferidos, e é composta uma imagem a partir desse interior da boca, da imagem sintetizada do ator alvo com os movimentos transferidos, e de uma imagem do vídeo original.

Finalmente, a solução foi avaliada através da comparação dos seus resultados com uma solução sofisticada do estado da arte, e também através de um questionário realizado por 23 pessoas. Neste questionário, os respondentes visualizaram vídeos com dobragens tradicionais e com dobragens criadas pela solução desenvolvida, e responderam a questões relacionadas com a qualidade de cada dobragem.

Os resultados obtidos revelaram que as dobragens criadas pela solução desenvolvida ainda não são melhores que as dobragens tradicionais, como era esperado. Mesmo assim, a solução desenvolvida pode ser considerada uma boa base de partida para obter resultados equivalentes às dobragens tradicionais e, eventualmente, ultrapassar essas dobragens.

7.1 Trabalho futuro

Existem vários passos a tomar para melhorar os resultados obtidos pela solução desenvolvida. A solução partilha de maior parte das limitações mencionadas no capítulo 3, como o resto do estado da arte, pelo que será necessário a melhoria contínua no sentido de reduzir e eliminar estas limitações. Dado o método de captura da geometria facial não ter sido desenvolvido especificamente para uma solução de captura e reprodução de movimentos faciais, no futuro será necessário implementar um método mais apropriado (como o de Thies et al. (2016)), ou adaptar o método utilizado. Uma melhoria na captura da geometria e dos movimentos faciais levaria diretamente a uma melhoria da transferência dos movimentos, e a uma melhoria das dobragens criadas.

Foram utilizadas 29 *blendshapes*, o que limita de certa forma os movimentos que podem ser capturados. No futuro, o número de *blendshapes* deveria ser aumentado, através da adaptação de *blendshapes* existentes noutros modelos 3D, ou através da criação manual de novas *blendshapes*. No entanto, como o algoritmo de captura não foi desenvolvido para um número elevado de *blendshapes*, será também necessário adaptá-lo.

Quanto à síntese do interior da boca, e à captura dos olhos, como já foi mencionado anteriormente, existem métodos de captura de modelos 3D de dentes e de olhos, que deverão ser implementados para melhorar as dobragens criadas pela solução.

Referências

Baltrusaitis, T., Robinson, P. & Morency, L.P., 2016. OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*.

Bregler, C., Covell, M. & Slaney, M., 1997. Video Rewrite. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, pp.353–360. Available at: http://dl.acm.org/citation.cfm?id=258734.258880.

Cao, C. et al., 2013. 3D Shape Regression for Real-time Facial Animation. *ACM Transactions on Graphics*, 32(4), p.41:1--41:10. Available at: http://dl.acm.org/citation.cfm?id=2461912.2462012.

Cao, C., Weng, Y., et al., 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3), pp.413–425. Available at: http://ieeexplore.ieee.org/document/6654137/.

Cao, C. et al., 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, 34(4), p.46:1-46:9. Available at: http://dl.acm.org/citation.cfm?doid=2809654.2766943.

Cao, C., Hou, Q. & Zhou, K., 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics*, 33(4), pp.1–10. Available at: http://dl.acm.org/citation.cfm?doid=2601097.2601204.

Chrysos, G.G. et al., 2017. A Comprehensive Performance Evaluation of Deformable Face Tracking «In-the-Wild». *International Journal of Computer Vision*. Available at: http://link.springer.com/10.1007/s11263-017-0999-5.

European Commission, 2012. *Europeans and their Languages*, Available at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf.

European Commission, 2006. *Europeans and their Languages*, Available at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf.

Garrido, P., Zollhöfer, M., Wu, C., et al., 2016. Corrective 3D reconstruction of lips from monocular video. *ACM Transactions on Graphics*, 35(6), pp.1–11. Available at: http://dl.acm.org/citation.cfm?doid=2980179.2982419 [Acedido Fevereiro 4, 2017].

Garrido, P. et al., 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics*, 32(6), pp.1–10. Available at: http://dl.acm.org/citation.cfm?doid=2508363.2508380 [Acedido Dezembro 6, 2016].

Garrido, P., Zollhöfer, M., Casas, D., et al., 2016. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Transactions on Graphics (TOG)*, 35(3), p.28. Available at: http://doi.acm.org/10.1145/XXXXXXX.YYYYYYY.

Garrido, P. et al., 2015. VDub: Modifying Face Video of Actors for Plausible Visual Alignment to

a Dubbed Audio Track. *Computer Graphics Forum*, 34(2), pp.193–204. Available at: http://doi.wiley.com/10.1111/cgf.12552 [Acedido Fevereiro 4, 2017].

Higginbotham, V., 1988. *Spanish film under Franco*, University of Texas Press.

Huang, H. et al., 2011. Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition. *ACM Transactions on Graphics*, 30(4), p.1.

Huber, P. et al., 2016. A Multiresolution 3D Morphable Face Model and Fitting Framework. *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, (February 2016), pp.79–86. Available at: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005669500790086.

Jin, X. & Tan, X., 2016. Face Alignment In-the-Wild: A Survey. , 6491. Available at: http://arxiv.org/abs/1608.04188.

Kazemi, V. & Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1867–1874.

Koen, P. et al., 2002. Fuzzy Front End: Effective Methods, Tools, and Techniques.

Koen, P. et al., 2001. Providing Clarity and a Common Language To the «Fuzzy Front End.» *Research Technology Management*, 44(2), pp.46–55.

Koolstra, C.M., Peeters, A.L. & Spinhof, H., 2002. The Pros and Cons of Dubbing and Subtitling. *European Journal of Communication*, 17(3), p.325. Available at: http://ejc.sagepub.com/cgi/content/abstract/17/3/325.

Malleson, C. et al., 2016. FaceDirector: Continuous control of facial performance in video. *Proceedings of the IEEE International Conference on Computer Vision*, 11–18–Dece, pp.3979–3987.

Matthias Niessner, 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral) - YouTube. Available at: https://www.youtube.com/watch?v=ohmajJTcpNk [Acedido Fevereiro 25, 2017].

McGurk, H. & MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, 264(December), pp.746–748.

Paysan, P. et al., 2009. A 3D Face Model for Pose and Illumination Invariant Face Recognition. *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, pp.296–301.

Presidência do Conselho, 1948. *Lei n.º 2027 de 18 de Fevereiro*, Available at: https://dre.pt/application/file/153175.

Shi, F. et al., 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics*, 33(6), pp.1–13. Available at: http://dl.acm.org/citation.cfm?doid=2661229.2661290.

Sumby, W.H. & Pollack, I., 1954. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2), pp.212–215. Available at: http://asa.scitation.org/doi/10.1121/1.1907309 [Acedido Fevereiro 5, 2017].

Sumner, R.W., 2005. *Mesh modification using deformation gradients*.

Sumner, R.W. & Popović, J., 2004. Deformation transfer for triangle meshes. *ACM Transactions on Graphics*, 23(3), p.399. Available at: http://portal.acm.org/citation.cfm?doid=1015706.1015736.

Thies, J. et al., 2015. Real-time Expression Transfer for Facial Reenactment. *SIGGRAPH Asia 2015*, 34(6), pp.1–14. Available at: http://people.mpi-inf.mpg.de/~mzollhoef/Papers/SGASIA2015_RR/page.html.

Thies, J., Zollhöfer, M. & Stamminger, M., 2016. Face2face: Real-time face capture and reenactment of rgb videos. Em *… Vision and Pattern …*. pp. 2387–2395. Available at: http://graphics.stanford.edu/~niessner/papers/2016/1facetoface/thies2016face.pdf [Acedido Novembro 11, 2016].

Tran, A.T. et al., 2016. Regressing Robust and Discriminative 3D Morphable Models with a very Deep Neural Network. *arXiv preprint arXiv:1612.04904*. Available at: http://arxiv.org/abs/1612.04904.

Valgaerts, L., Wu, C. & Seidel, H., 2012. Lightweight Binocular Facial Performance Capture under Uncontrolled Lighting.

Vlasic, D. et al., 2005. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3), p.426.

Wagner, P., 2011. Local Binary Patterns. , p.1. Available at: https://bytefish.de/blog/local_binary_patterns/ [Acedido Outubro 16, 2017].

Wood, E. et al., 2016. A 3D morphable eye region model for gaze estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS, pp.297–313.

Woodall, T., 2003. Conceptualising 'Value for the Customer': An Attributional, Structural and Dispositional Analysis. *Academy of Marketing Science Review*, 12(5), pp.1–42.

Wu, C. et al., 2016. Model-based teeth reconstruction. *ACM Transactions on Graphics*, 35(6), pp.1–13. Available at: http://dl.acm.org/citation.cfm?doid=2980179.2980233 [Acedido Dezembro 6, 2016].

Anexos

Tabela 4 – Respostas ao vídeo 1

	Quão natural lhe pareceu a dobragem?	Quão natural lhe pareceu o movimento da boca neste vídeo?	Quão confortável se sentiu ao ver este vídeo?	Como avaliaria a sincronização entre o áudio e a imagem?	Como avaliaria a sua compreensão do que foi dito neste vídeo?
Respondente 1	3	2	3	4	4
Respondente 2	2	4	2	2	5
Respondente 3	4	3	3	4	3
Respondente 4	5	4	4	4	5
Respondente 5	4	4	1	5	5
Respondente 6	3	3	3	4	5
Respondente 7	3	4	2	4	5
Respondente 8	3	2	3	4	5
Respondente 9	3	2	2	2	4
Respondente 10	4	3	4	3	4
Respondente 11	4	4	3	4	3
Respondente 12	3	2	3	2	4
Respondente 13	2	1	2	1	3
Respondente 14	4	4	4	5	5
Respondente 15	4	4	3	5	5
Respondente 16	4	3	4	5	5
Respondente 17	3	4	4	3	4
Respondente 18	3	3	3	3	5
Respondente 19	4	5	4	4	5
Respondente 20	3	2	2	5	5
Respondente 21	5	4	3	5	5
Respondente 22	3	2	3	3	4
Respondente 23	5	4	5	5	5

Tabela 5 – Respostas ao vídeo 2

	Quão natural lhe pareceu a dobragem?	Quão natural lhe pareceu o movimento da boca neste vídeo?	Quão confortável se sentiu ao ver este vídeo?	Como avaliaria a sincronização entre o áudio e a imagem?	Como avaliaria a sua compreensão do que foi dito neste vídeo?
Respondente 1	3	1	2	3	4
Respondente 2	1	1	2	1	5
Respondente 3	2	2	2	2	4
Respondente 4	3	2	3	3	5
Respondente 5	1	1	1	5	5
Respondente 6	1	2	2	2	5
Respondente 7	1	1	1	3	5
Respondente 8	2	1	2	4	5
Respondente 9	1	1	1	1	4
Respondente 10	1	1	1	1	1
Respondente 11	1	1	1	1	2
Respondente 12	1	1	2	2	2
Respondente 13	2	1	2	3	4
Respondente 14	2	1	2	1	3
Respondente 15	1	1	1	1	5
Respondente 16	1	2	3	4	3
Respondente 17	3	4	4	3	3
Respondente 18	1	1	3	3	5
Respondente 19	1	1	1	3	3
Respondente 20	2	1	2	4	5
Respondente 21	1	1	2	2	4
Respondente 22	1	1	1	3	4
Respondente 23	1	1	3	1	5

Tabela 6 – Respostas ao vídeo 3

	Quão natural lhe pareceu a dobragem?	Quão natural lhe pareceu o movimento da boca neste vídeo?	Quão confortável se sentiu ao ver este vídeo?	Como avaliaria a sincronização entre o áudio e a imagem?	Como avaliaria a sua compreensão do que foi dito neste vídeo?
Respondente 1	2	2	2	3	4
Respondente 2	1	5	2	2	5
Respondente 3	3	3	3	4	4
Respondente 4	5	4	4	4	5
Respondente 5	4	3	5	2	5
Respondente 6	3	4	3	3	5
Respondente 7	5	5	4	5	5
Respondente 8	3	3	3	4	5
Respondente 9	2	2	1	4	4
Respondente 10	3	2	2	2	3
Respondente 11	3	3	2	3	4
Respondente 12	3	3	3	4	5
Respondente 13	1	1	2	1	4
Respondente 14	5	4	5	4	5
Respondente 15	4	4	3	4	5
Respondente 16	2	2	3	1	4
Respondente 17	5	3	4	4	3
Respondente 18	3	3	3	3	5
Respondente 19	5	5	5	5	5
Respondente 20	5	5	4	5	5
Respondente 21	4	4	4	5	5
Respondente 22	2	5	4	5	4
Respondente 23	3	5	4	3	5

Tabela 7 – Respostas ao vídeo 4

	Quão natural lhe pareceu a dobragem?	Quão natural lhe pareceu o movimento da boca neste vídeo?	Quão confortável se sentiu ao ver este vídeo?	Como avaliaria a sincronização entre o áudio e a imagem?	Como avaliaria a sua compreensão do que foi dito neste vídeo?
Respondente 1	3	3	3	4	4
Respondente 2	1	2	2	2	5
Respondente 3	2	3	2	2	4
Respondente 4	2	2	2	4	5
Respondente 5	3	2	2	3	5
Respondente 6	2	2	2	3	4
Respondente 7	2	1	2	2	5
Respondente 8	3	3	3	4	5
Respondente 9	2	2	1	2	4
Respondente 10	3	2	2	2	2
Respondente 11	1	1	1	3	1
Respondente 12	1	1	2	2	4
Respondente 13	2	3	3	3	4
Respondente 14	3	3	3	2	5
Respondente 15	1	1	1	1	5
Respondente 16	2	2	3	3	4
Respondente 17	4	3	3	4	4
Respondente 18	2	2	3	3	5
Respondente 19	3	3	3	4	5
Respondente 20	3	2	3	4	5
Respondente 21	1	3	1	3	4
Respondente 22	1	1	2	4	4
Respondente 23	2	2	4	3	5