Como citar este artículo:T. Oliveira, N. Escudeiro, P. Escudeiro, E. Rocha and F. M. Barbosa, "The VirtualSign Channel for the Communication Between Deaf and Hearing Users," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 14, no. 4, pp. 188-195, Nov. 2019, doi: 10.1109/RITA.2019.2952270.

O VirtualSign como Canal de Comunicação entre Utilizadores Surdos e Ouvintes

Tiago Oliveira, Nuno Escudeiro, Paula Escudeiro, Emanuel Rocha, Fernando Maciel Barbos

***Title*—The VirtualSign Channel for the Communication Between Deaf and Hearing Users**

***Abstract*— Deaf students, who use sign language as their mother language, continuously experience difficulties to communicate with non-deaf in their daily lives. This is a severe handicap in education settings seriously jeopardizing deaf people chances to progress in their professional career. Deaf people's comprehension of texts is limited due to grammar differences between sign and oral languages. There is a need to improve the communication between deaf and non-deaf and to support deaf students in environments where they are unable to be accompanied by sign interpreters. This article details the improvements and current structure of the VirtualSign platform, a bidirectional sign language to text translation tool in development since 2015. The platform has two main components, sign to text and text to sign, that are both described. Translation from text to sign relies on a 3D avatar. Translation from sign to text relies on a set of data gloves and Kinect. In this paper we discuss the relevance of different types of data gloves. VirtualSign is being developed in cooperation with the deaf communities from six different European countries and Brazil. This solution to support deaf students in educational settings has received positive feedback on several tests and pilot experiments. Some planned improvements and future functionalities for the tool are also mentioned and detailed.**

***Index terms*—Sign Language, Sign Language Recognition, Translation, Deaf, Hearing Disabilities, Accessibility, Inclusion**

I. Introduction

ATUALMENTE, cerca de 15-26% da população mundial sofre de algum tipo de deficiência auditiva, com cerca de 10 000 pessoas a perderem a audição antes de terem aprendido qualquer língua [1], o que faz com que adotem a língua gestual do seu país como principal forma de comunicação. De acordo com dados recentes Federação Mundial de Surdos, mais de 70 milhões de pessoas sabem língua gestual, e existem mais de 300 línguas gestuais diferentes por todo o mundo [2]. Ao contrário da ideia comum, nem toda a gente que utiliza língua gestual consegue ler e interpretar texto escrito da mesma forma de pessoas que usam língua oral, isto acontece devido às diferenças na estrutura das frases e gramática entre os dois tipos de linguagem [3]. Com o aumento nas duas últimas décadas do número de surdos a frequentar o ensino superior [4], [5], esta dificuldade de interpretação pode tornar-se um problema para a comunicação entre professor e aluno, ou até mesmo entre alunos. [6]. Apesar de alguns países já investirem no assunto e terem começado a adaptar as instituições de ensino a esta realidade, [7]visto que as mudanças propostas pela política oficial para a educação de surdos em todos os níveis não estão ainda efetivamente implantadas nem as instituições educacionais preparadas para responder às necessidades desses alunos. Essa preocupação conduziu à realização da presente pesquisa, que teve o objetivo de analisar, por meio de entrevistas individuais, os dizeres de alunos surdos e seus professores universitários a respeito dos obstáculos e das possibilidades que o surdo encontra no seu cotidiano de estudo. Os sujeitos foram quatro alunos e seis professores de duas universidades. Com os alunos foram realizadas entrevistas presenciais de caráter semi-estruturado, em LIBRAS, que foram videogravadas e posteriormente traduzidas para o Português. Com os professores as entrevistas aconteceram por meio da internet, utilizando e-mail e Messenger com vídeo conferência. Na fase de análises, os depoimentos dos dois grupos foram organizados em quatro unidades temáticas: 1, [8], ainda há muito trabalho a fazer para tornar o processo de aprendizagem tão fluído para os alunos surdos como é para os restantes. Ter um intérprete de língua gestual na sala de aula é uma das soluções mais comuns, mas só funciona em sala de aula e não ajuda o aluno em ambientes fora da instituição de ensino quando o mesmo está sozinho, p. ex. estudar slides PowerPoint ou documentos escritos em casa em preparação para um exame.

Esperamos que um tradutor automático bidirecional entre texto escrito e língua gestual possa ajudar na comunicação entre estudantes surdos e ouvintes e professores, de forma a quebrar as barreiras de comunicação e melhorar a experiência de ensino/aprendizagem de todos os intervenientes.

Este artigo é uma versão expandida do artigo publicado na EDUCON 2019 [9], e inclui um estado da arte revisto, detalhes adicionais acerca do processo de tradução de gesto para texto, e a descrição de um plano experimental para

T. Oliveira é investigador no Instituto Superior de Engenharia do Porto (e-mail: taqol@isep.ipp.pt, ORCID: 0000-0002-7063-686X)
N. Escudeiro é professor no Instituto Superior de Engenharia do Porto (e-mail: nps@isep.ipp.pt, ORCID: 0000-0003-3940-3846)
P. Escudeiro é professora no Instituto Superior de Engenharia do Porto (e-mail: pmo@isep.ipp.pt, ORCID: 0000-0003-2528-572X)
E. Rocha é investigador no Instituto Superior de Engenharia do Porto (e-mail: evrev@isep.ipp.pt, ORCID: 0000-0002-3031-1884)
F. Maciel Barbosa é professor na Faculdade de Engenharia da Universidade do Porto (e-mail: fmb@fe.up.pt, ORCID: 0000-0003-1065-2371)

avaliar sistemas de tradução utilizando diferentes modelos de luvas de dados.

II. Sign Language

Como mencionado anteriormente, existe a perceção comum de que, embora os alunos surdos usem língua gestual como língua principal, estes conseguem ler textos fluentemente na língua oral dos seus países, porque a sua visão não é prejudicada. Essa perceção não corresponde à realidade. Uma língua gestual é uma língua por si só, reconhecida como língua oficial do país na maioria dos países europeus. Como idioma, possui um conjunto de regras gramaticais e estrutura de frases específica. Essas diferenças, aliada ao facto dos surdos aprenderem a língua oral/escrita do seu país como segunda língua, são a principal razão pela qual eles não conseguem ler com fluência textos escritos. Por exemplo, a frase "vou para casa" em português pode ser "casa ir" em língua gestual; para uma pessoa cuja língua principal é a gestual, todas as outras palavras da frase são vistas como ruído que dificulta a compreensão [3].

As línguas gestuais em todo o mundo são diferentes entre si, assim como as línguas escritas, mas partilham os mesmos três componentes principais: configuração de mão, expressão facial e movimento do corpo. A configuração de mão é considerada a parte principal do gesto e é composta pela posição e flexão dos dedos na mão. A maioria das configurações, quando feitas isoladamente, representam uma letra ou um número da língua gestual (como mostrado na Figura 1), mas também existem configurações específicas para gestos.

A expressão facial é outra parte importante da língua gestual, pois são a única maneira de associar emoção a uma determinada frase e são essenciais para saber se a frase é uma pergunta ou afirmação, p. ex. usando olhares duvidosos ou diretivos. Em alguns idiomas, o componente de expressão facial também inclui movimento da boca ao som da palavra, como se fosse falado, mas esse não é um caso geral. O movimento corporal é o gesto em si, e é composto pelo movimento não apenas das mãos, mas de todo o corpo. Esse movimento pode ser realizado com uma mão ou ambas, mas a parte principal do movimento será realizada pela mão dominante, seja ele canhoto ou destro.

Como o objetivo deste trabalho é a tradução em tempo real, todos esses componentes foram tidos em conta: as regras gramaticais para tradução de texto para gesto, e os três componentes dos gestos para a tradução gesto para texto.

Fig. 1. O alfabeto da Língua Gestual Portuguesa (configurações de mão) [3]

III. Objectivos

O projeto VirtualSign já foi discutido anteriormente [5], [6] mas tem vindo a evoluir e melhorar ao longo dos anos para acompanhar as soluções e tecnologias mais recentes. O VirtualSign é um produto que consiste num tradutor bidirecional entre texto e língua gestual. É possível introduzir texto e mostrar um avatar 3D a traduzir em língua gestual de um país específico, ou ter um utilizador a usar língua gestual em frente a uma câmara e apresentar a tradução em texto escrito. A ideia da ferramenta surgiu em sala de aula quando um professor de engenharia descobriu que a comunicação com um aluno com deficiência auditiva apenas através dos materiais da aula era complicada, e não havia funcionários da universidade preparados para a situação.

O projeto I-ACE[1] visa melhorar a eficácia da comunicação entre surdos e ouvintes em ambiente educacional através de três aplicações principais: (1) durante as aulas, apresentando o avatar 3D, traduzindo o texto escrito para língua gestual; (2) permitindo ao aluno fazer uma pergunta em língua gestual, traduzi-la para a linguagem oral correspondente e apresentá-la em texto escrito ao professor e (3) dar ao aluno a possibilidade de traduzir qualquer material escrito fora da sala de aula por meio de um tradutor on-line / móvel.

IV. Estado da Arte

Nos últimos anos, de acordo com as estatísticas da Thomson Reuters Web of Science, houve um aumento significativo de publicações e pesquisa na área de reconhecimento de língua gestual nas áreas de engenharia eletrónica e informática, principalmente nos EUA, Índia e China [12].

Muitas tentativas de reconhecimento de língua gestual limitam-se ao reconhecimento de *fingerspelling*, ou seja, o reconhecimento de uma única letra ou número de cada vez, com base na posição dos dedos da mão (conhecida como "configuração de mão»). Além disso, as tentativas de traduzir palavras reais acabam sendo com vocabulário reduzido. Tentativas recentes como o sistema de Taskiran, Killioglu e Kahraman [13], que usam um modelo baseado em redes neuronais convolucionais (CNN), treinado com um conjunto de imagens cortadas para diferentes configurações de mão da Língua Gestual Americana. As CNN são uma técnica de *deep learning*, com resultados comprovados no processamento de imagens [14], e, por fazer uso dos processadores gráficos dos computadores modernos e não exigir quase nenhum pré-processamento, consegue ser mais rápido que outras técnicas de classificação nessa área. Nos testes em tempo real, o sistema criado por eles conseguiu alcançar um resultado de 98,05% de precisão.

Ainda no tópico das CNN, Kumar et al. [15] usam essa técnica para analisar imagens criadas por mapas de deslocamento angular de articulações enquanto o utilizador faz os gestos. Esses mapas são uma combinação de medidas angulares com mapas de distâncias conjuntas que essencialmente capturam os dados do esqueleto do utilizador em imagens 3D que podem ser processadas pela rede. Com o seu próprio *dataset* de 200 palavras em Língua Gestual

1 I-ACE - International Assisted Communication for Education, 2016-1-PT01-KA201-022812, foi cofinanciado pelo programa Erasmus+ entre 2016 e 2018 para promover a internacionalização da plataforma VirtualSign, com vista a estender a tradução para seis línguas gestuais Europeias e apoiar o sistema educativo.

Indiana, conseguem atingir 90% de precisão e melhorar os resultados anteriores que utilizavam a mesma técnica, mas ainda assim não é um sistema em tempo real.

Liao et al. [16] usam um método dinâmico de reconhecimento de língua gestual usando CNN. Em vez de processar a imagem completa de cada vídeo, é feita a localização da mão primeiro, extraída, e os restantes *frames* são analisados em função disso, um processo que reduz o tempo e a complexidade dos cálculos na rede. No final conseguem atingir uma média de 88% nos dois *datasets* que testaram, mas ainda é apenas uma tradução em vídeo e não em tempo real.

Ainda no campo do reconhecimento da língua gestual através de vídeos, o trabalho recente de Yuan et al. [17] traz um futuro promissor à área, apresentando o maior dataset de vídeo de língua gestual até hoje, com mais de 50.000 vídeos de Língua Gestual Chinesa, para ajudar a aprimorar as técnicas de aprendizagem automática sobre o assunto.

Huang et al. [18]thus the automatic translation of a sign language is meaningful and important. Currently, there are two sub-problems in Sign Language Recognition (SLR propõem uma solução interessante de tradução em tempo real, baseada em processamento de vídeo. São estudados dois *streams* diferentes de cada vídeo em língua gestual, um com o utilizador e o movimento completo, e outro focado apenas na posição da mão em cada *frame*. O resultado é transmitido para uma CNN para processamento de vídeo e conseguem atingir uma precisão de 82,7% com um vocabulário de 178 palavras. Mas, embora seja considerada uma tradução em tempo real, fazem uso de gestos específicos para assinalar o início e o final de cada frase, algo que é impraticável para uso em ambiente real.

Rao et al. [19] desenvolveram um sistema contínuo de tradução de língua gestual usando a câmara frontal de um smartphone para analisar gestos de Língua Gestual Indiana feitos apenas com uma das mãos. O sistema utiliza técnicas de processamento de vídeo para determinar padrões com base no contorno da mão na imagem em cada *frame*. O melhor resultado é 90,58%, mas o vocabulário é composto por apenas 18 palavras e todas tiveram de ser executadas em ambiente com iluminação controlada e fundo simples ao gravar o vídeo, o que é um grande obstáculo em termos práticos.

Até agora, nenhuma dessas soluções possuía tecnologia em comum com a nossa proposta. Porém, a solução de Dong, Leu e Yin [20] visa reconhecer o alfabeto de Língua Gestual Americana usando o mesmo sensor de imagem que o VirtualSign, o sensor Kinect. Este produto da Microsoft é amplamente utilizado e consiste numa câmara com sensores de profundidade e cor, combinados com microfones para captura de movimento 3D, reconhecimento facial e de voz [21]. A solução deles usa o Kinect não para obter o movimento do corpo, mas apenas a configuração das mãos e determiná-la com um classificador *Random Forest*, mas só consegue traduzir letra por letra os 24 sinais presentes no alfabeto da Língua Gestual Americana. Ainda assim, alcançou 92% de precisão, o que é muito bom em comparação com outras soluções similares.

Para resumir, de todas as soluções recentes analisadas, a maioria não é em tempo real, e as que são traduzem apenas letra a letra. As soluções que tentam ir mais além funcionam apenas com uma base de dados muito pequena. A nossa solução visa melhorar isso, com um alfabeto completo para cada língua e um vocabulário com 500 palavras para os idiomas menos completos e mais de 2000 para o nosso idioma original: português, todas compostas por configuração de mão e movimento.

V. VirtualSign

A plataforma VirtualSign é composta por duas áreas principais: tradução de texto para gesto e tradução de gesto para texto. Cada uma dessas áreas deu origem a uma aplicação, um tradutor. Uma terceira ferramenta, o VirtualSign Studio Online (VSSO), dá suporte a todas as traduções, é um configurador on-line de língua gestual.

O tradutor de texto para gesto (TG) e a aplicação VSSO são ferramentas on-line desenvolvidas no Unity e incorporadas com WebGL para aumentar a compatibilidade com vários navegadores e permitir que os utilizadores as usem em qualquer lugar. O tradutor de gesto para texto é uma aplicação *desktop* que faz uso de luvas de dados e um sensor Kinect. Embora a base para ambas as aplicações seja a tradução de e para o texto, atualizamos nosso sistema com tecnologia de texto para voz para futuras implementações.

A. Tradução Texto para Gesto

O tradutor TG é usado como os tradicionais tradutores online, onde o utilizador escreve o texto numa caixa em branco e vê instantaneamente o resultado noutro idioma. Neste caso, o utilizador pode selecionar entre seis diferentes línguas gestuais europeias implementadas até o momento: portuguesa, britânica, cipriota, alemã, grega e eslovena. O resultado é o nosso avatar 3D a fazer os gestos correspondentes ao texto escrito. O texto introduzido é passados pelas regras gramaticais da língua especificada e alterado de acordo (ordem das palavras e tempos verbais), com a frase resultante a ser pesquisada na nossa base de dados de gestos e os movimentos são transmitidos ao avatar para executar. É um processo quase instantâneo e detalhes como o processamento da sintaxe são escondidos do utilizador, o que o torna muito intuitivo.

B. Construir o Léxico (Configurador VSSO)

O configurador VSSO é onde é construído o vocabulário de gestos para todo o projeto, com a participação de surdos e intérpretes que adicionam e validam palavras com esta ferramenta, para que possam ser usadas nos tradutores. A interface do configurador é semelhante às ferramentas de animação 3D como o Maya ou o Blender. O avatar 3D pode ser manipulado com o rato em estilo *drag and drop*, semelhante a outras aplicações de animação 3D. Ambos os braços, corpo e cabeça podem ser manipulados com o rato para obter diferentes graus de movimento, e a expressão facial e configuração de mão são alteradas através de listas com opções pré-definidas. Cada língua tem o seu próprio conjunto de configurações de mão diferentes, mas as expressões faciais são seis comuns: neutro, feliz, triste, surpreendido, inquisitivo e zangado. Os movimentos são criados registando os diferentes *keyframes* que compõem o movimento, como mostra a Figura 2. Ao ter esses *keyframes*, o software faz a interpolação do movimento intermédio, criando o gesto final. Para facilitar a criação e consumir menos tempo, a ferramenta

Fig. 2. O gesto para "Bom Dia" em Língua Gestual Portuguesa, decomposto em *keyframes* [10

permite copiar *keyframes* anteriores, para movimentos que exigem repetição, e importar *keyframes* de outros gestos na base de dados, para gestos que sejam semelhantes.

Para a configuração das palavras, consideramos dois tipos de utilizadores: editores e validadores. Cada utilizador tem um login diferente que o define como um tipo ou outro, com base na sua experiência em língua gestual. O editor pode ser qualquer utilizador comum, com ou sem experiência em língua gestual, pois é possível imitar movimentos de vídeos e dicionários de língua gestual na ferramenta, manipulando o avatar e guardando o gesto na base de dados. Esses gestos ficam marcados como "não validados" e não são utilizados nos tradutores até que sejam aprovados por um validador. Os utilizadores considerados validadores devem ser especialistas em língua gestual, seja por usarem uma língua gestual como primeira língua, ou por serem intérpretes certificados, e são definidos manualmente como tal pela equipa de desenvolvimento e, além de poderem adicionar palavras à ferramenta, são responsáveis por validar os gestos criados pelos editores. Para fazer isso, a ferramenta possui um sistema de mensagens no qual o validador pode fornecer feedback ao editor responsável por um gesto específico, caso os movimentos precisem de algumas correções, ou podem apenas validá-lo ou excluí-lo da base de dados, caso não se encaixe nos padrões.

Além desse processo de configuração, o VSSO possui também uma ferramenta de contexto, no qual o utilizador pode selecionar um contexto para uma palavra através de uma lista na aplicação. Essa lista é preenchida com conteúdo de um dicionário on-line em tempo real, considerando a língua gestual selecionada e a palavra que o utilizador deseja inserir. Esse contexto será usado posteriormente no processo de tradução para desambiguar entre palavras semelhantes ou contexto confuso quando estiver a ser traduzida uma frase ou texto completo.

C. Tradução Gesto para Texto

A tradução de gesto para texto requer dois componentes diferentes. Uma aplicação de captura de mão, usada para criar *datasets* para classificação de configurações de mão, e o tradutor em si. Esses componentes usam dois dispositivos externos para recolher dados do utilizador durante a tradução: um par de luvas de dados e um sensor Microsoft Kinect.

Neste momento estamos também a trabalhar no processamento de imagens com webcams normais, semelhantes às experiências referidas anteriormente no estado da arte. Os resultados obtidos até agora não são tão satisfatórios quanto os que obtemos com nossa abordagem atual devido ao tamanho do vocabulário, à quantidade de idiomas e à potência do hardware à nossa disposição. O nosso objetivo é disponibilizar amplamente a aplicação em locais onde os computadores podem não ser modernos ou muito potentes. Portanto, precisamos de uma solução que possa ser executada em hardware comum. No entanto, temos sempre como objetivo futuro reduzir o número de dispositivos externos necessários para a tradução.

Atualmente estamos a trabalhar com dois modelos diferentes de luvas de dados, um construído por uma empresa externa, a Fifth Dimension Technologies, e outro desenvolvido no Brasil, em cooperação connosco. Os modelos usam sensores diferentes para atingir o mesmo objetivo, e as suas características estão detalhadas na secção VI. Usamos esses modelos para capturar a configuração e a rotação das mãos, combinadas com um sensor Kinect, versão 2, responsável por capturar a localização e o movimento das mãos através do objeto *body* fornecido pela Microsoft no seu SDK.

A aplicação de captura de mão é um programa de *desktop* simples para capturar os *inputs* das luvas de dados. Esse *dataset* é usado para treinar um classificador para configurações de mão. O utilizador usa a luva, seleciona a língua gestual desejada e inicia o processo de calibração. Esse processo requer que o utilizador execute um conjunto pré-definido de configurações de mão, enquanto as luvas são calibradas automaticamente para garantir que todos os *datasets* sejam criados de igual forma. Após o processo de calibração, a captura começa. Nesta fase, a aplicação mostra uma imagem de uma configuração de mão, o utilizador coloca a mão como apresentado na imagem e regista-a pressionando uma tecla. O objetivo é repetir a configuração com a maior precisão possível um determinado número de vezes (10, 20, …), passando para a próxima e repetindo o processo, de forma a criar um *dataset* para as configurações de mão usadas na língua gestual. Este *dataset* será usado para treinar um classificador a ser usado pelo tradutor. A aplicação suporta qualquer modelo de classificação, incluindo, por exemplo, *K-Nearest Neighbors* [22] e *Convolutional Neural Networks*.

A segunda aplicação é o tradutor de gesto para texto. É mais complexo que o tradutor de texto para gesto, pois existem mais fatores que têm de ser observados para uma tradução precisa. Atualmente ainda não há processamento automático para as expressões faciais. Esta é uma das melhorias planeadas para o futuro próximo.

A tradução é feita carregando todo o léxico construído com a ferramenta VSSO, ou seja, carregando todas as palavras com os seus respetivos *keyframes* compostos por coordenadas da mão, rotação, configuração de mão e movimento do corpo. Com essas informações, o sistema está pronto para gerar um grafo de conversão, uma espécie de estrutura em árvore em que cada nó representa uma mudança no estado do movimento, e é construído globalmente. O processo básico de tradução é capturar o primeiro movimento, procurá-lo no primeiro nível do grafo, capturar o segundo movimento e tentar encontrá-lo num dos filhos do primeiro nó, ou começar de novo se não for encontrado. O processo continua até que seja alcançada uma folha da árvore e a palavra é apresentada.

Para conseguir isso, definimos as mudanças no movimento como eventos que ficam associados a um tipo de nó, e, nesta primeira versão, consideramos três tipos. O primeiro tipo de evento ocorre quando a configuração de mão é alterada, e esse tipo de nó é criado quando, de um *keyframe* para outro, a configuração de mão do avatar 3D é alterada; O segundo evento é quando é detetada uma rotação da mão, quando a mão do avatar roda além de uma certa margem pré-definida;

e o terceiro evento é despoletado pelo movimento do corpo, especificamente o movimento da mão. Este último funciona através da análise do movimento, se este é linear entre os *keyframes* e ativa quando o movimento se mantém contínuo, mas deixa de ser linear, e isso produz resultados como "mão esquerda movida para cima para a esquerda" ou "mão direita movida para a esquerda para baixo". Para resumir, um exemplo simples como "Bom dia" em Língua Gestual Portuguesa, se considerarmos apenas a mão dominante, produziria "B, esquerda cima, S", onde B e S correspondem a configurações de mão. Uma limitação do sistema atual é que alguns *keyframes* para certas palavras têm mais do que uma alteração, p. ex. um utilizador pode mover a mão para cima e alterar a configuração de mão também, mas estamos a estudar uma possível implementação com probabilidades aplicadas aos nós do grafo para resolver esse problema.

O tradutor de gesto para texto constrói o grafo no momento em que é iniciado e de forma invisível para o utilizador, ao qual é apresentada a interface, uma única janela com um menu e o *feed* da câmara onde o utilizador se consegue ver a si mesmo e os ícones que representam o estado de ligação das luvas. O utilizador deve então escolher uma língua gestual e segue para o processo de calibração (o mesmo que na aplicação de captura de mão). Uma vez concluído, o processo de tradução pode começar com apenas um toque de um botão e é contínuo e em tempo real, o utilizador pode executar o número de gestos que quiser e ver a tradução abaixo da imagem da câmara no ecrã.

A aplicação observa o utilizador para perceber quando ocorre um evento, os mesmos eventos usados na criação do grafo a partir do léxico. A Figura 3 mostra um resumo dos eventos e as condições para a sua deteção na aplicação de tradução de sinal para texto.

O sensor Kinect e as luvas de dados monitorizam todo o corpo em tempo real, procurando ativamente condições para os eventos. O Kinect observa os movimentos das mãos e captura o ponto de partida e o movimento. Quando o movimento deixa de ser linear, o sensor ativa o evento e analisa a direção do movimento no espaço 3D, depois continua a observar as duas mãos para novos movimentos. As luvas estão a trabalhar nos outros dois eventos. Para o evento de rotação da mão, o giroscópio lê ativamente os dados de ambas as luvas e, uma vez que ultrapassem um certo limite, o evento é ativado e passa as informações da rotação (*pitch*, *roll* ou *yaw*; positivo ou negativo) para o tradutor. O último evento e, sem dúvida, o mais complexo dos três é a configuração de mão. Além de exigir captura dos sensores das luvas em tempo real, exige também um classificador a trabalhar constantemente em segundo plano, a determinar qual a configuração de mão que o usuário está a fazer para a língua selecionada. Os classificadores são os

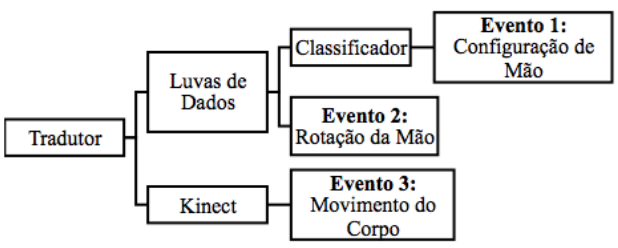


Fig. 3. Estrutura do tradutor de gesto para texto e os eventos

mesmos usados na aplicação de captura de mão, KNN ou CNN. Para esta primeira versão, estamos a considerar apenas o primeiro resultado do classificador, mas estamos a planear uma versão atualizada, que tenha em consideração múltiplas possibilidades, probabilidade associada aos eventos, e desambiguação recorrendo ao contexto.

O tradutor que trabalha em segundo plano está a receber os eventos detetados sequencialmente, faz uma leitura e percorre o grafo do vocabulário em tempo real, mostrando o resultado de cada vez que atinge uma folha, ou seja, uma palavra na língua oral que corresponde à sequência de gestos do utilizador. Quando isso acontece, o tradutor volta ao nó raiz e o processo inicia novamente.

1. *Modos de Funcionamento*

Outra característica que foi incorporada nesta ferramenta é o modo de *fingerspelling*. Na língua gestual, os nomes e algumas outras palavras que ainda não têm um gesto correspondente são soletrados, usando as configurações de mão do alfabeto da língua correspondente em sequência para construir a palavra. Por esse motivo, criamos dois modos no tradutor, o "modo tradução", para a tradução normal descrita nos parágrafos anteriores, e o "modo soletrar". Este último é ativado automaticamente quando o utilizador executa três configurações de mão que correspondam a letras de forma sequencial e sem mover significativamente a mão em qualquer direção. O modo de tradução é reativado assim que o utilizador volte a mover a mão.

2. *Add-On PowerPoint*

O primeiro produto resultante desta arquitetura chega na forma de um complemento do PowerPoint para tradução de texto para gesto. É uma ferramenta simples em que o utilizador, ao criar uma apresentação PowerPoint, pode selecionar texto de um slide, escolher uma língua gestual e obter instantaneamente um vídeo do avatar 3D a traduzir esse texto. Embora possa ser usado por qualquer pessoa, foi desenvolvido a pensar em professores e alunos. Os professores podem criar uma apresentação para as aulas e obter instantaneamente a tradução em vídeo para cada slide, permitindo assim que os alunos surdos acompanhem a aula. A vantagem para os alunos é poderem abrir qualquer PowerPoint em casa, das aulas ou de algum recurso on-line para estudar, e obter a tradução para uma língua gestual à sua escolha.

VI. Luvas de Dados

Como referido anteriormente, o VirtualSign requer dois dispositivos - um par de luvas de dados e um sensor de movimento Kinect - para obter os *inputs* usados no processo de tradução de gesto para texto em tempo real. Esses *inputs* alimentam o nosso modelo de tradução, através dos componentes da linguagem de sinais já descritos: configuração de mão, rotação e movimento. O Kinect fornece informações acerca do movimento das mãos e do corpo em geral. As luvas de dados são uma forma precisa de capturar a forma e a rotação das mãos, e eliminam várias dificuldades que existem noutros tipos de interfaces de gesto, como o Leap Motion, bem como de soluções com base em processamento de imagens. Não são influenciadas por "áreas sombra" que acontecem

quando as mãos se sobrepõem, não estão dependentes das condições de iluminação, e os *inputs* gerados não requerem pré-processamento pesado, tornando-as adequadas para uso em tempo real. A principal desvantagem está no facto de serem intrusivos na maneira como devem ser usados pelo utilizador, além disso, os modelos atualmente disponíveis são geralmente caros e exigem *drivers* específicos, bem como requisitos extra de configuração e ajustes.

Estamos atualmente a trabalhar com dois tipos de luvas de dados, 5DT e IFG.

A. Luvas de Dados 5DT

As luvas de dados usadas inicialmente no VirtualSign foram as 5DT Ultra 14 [23], fabricadas pela Fifth Dimension Technologies. Vistas de fora, são muito semelhantes a uma luva comum, mas dentro do tecido existem 14 sensores de flexão alinhados com as articulações dos dedos. Esse número de sensores ajudou a reduzir problemas já documentados provocados por diferentes tamanhos de mão ao executar a mesma configuração [24]. Esse modelo permitiu-nos criar a base para a versão atual do tradutor e conseguimos atualizar o processo usando dois modelos novos e mais recentes.

O primeiro dos novos modelos de luva de dados é uma versão melhorada das luvas originais: 5DT Ultra 17, desenvolvidas pela mesma empresa. O *design* do modelo é praticamente igual, mas foi adicionada uma caixa de pequena dimensão em cima da luva, que contém um giroscópio e um acelerómetro, que usamos para determinar a rotação da mão, algo que era impossível com o primeiro modelo.

B. Luvas de Dados IFG

O novo modelo que estamos atualmente a implementar é um par de luvas ainda sem nome desenvolvidas pelo Instituto Federal de Goiás, uma universidade situada em Goiânia, no Brasil, em colaboração connosco. Estas luvas foram desenhadas sem sensores de flexão e são compostas somente por 9 sensores contendo um acelerómetro e um giroscópio cada, 1 sensor por cada dedo, 1 no centro da luva, 1 junto ao pulso, e 2 no fundo do dedo indicador e do anelar, respetivamente.

Este modelo tem a vantagem de ser desenvolvido pelos nossos parceiros, pelo que pode ser alterado e melhorado de acordo com o nosso *feedback* durante os testes. É também uma alternativa mais barata, o que a vai tornar mais abrangente quando for lançado o produto final. O aumento do número de giroscópios e acelerómetros abre a possibilidade de desenvolver um tradutor sem recurso ao Kinect, mas a remoção dos sensores de flexão também traz desvantagens, nomeadamente o pré-processamento extra que é necessário para obter a configuração de mão correta.

C. Performance das Luvas de Dados

Os sensores de flexão são úteis porque fornecem dados constantes do sensor para cada configuração de mão, independentemente da orientação da mão, ou seja, independentemente da direção para onde a palma da mão está virada. Enquanto que os *inputs* gerados pelos múltiplos giroscópios e acelerómetros da luva IFG mudam de cada vez que a palma da mão fica virada para uma direção diferente. Por um lado, esse recurso dificulta a previsão da configuração da mão correta, pois esta não depende da orientação da mão, mas ao mesmo tempo abre caminho para um sistema que se baseie completamente apenas nas luvas de dados. A diversidade de sensores permite-nos testar a sua influência na tradução e planear alterações futuras para melhorar o desempenho da tradução de gesto para texto.

Para poder comparar os resultados do nosso sistema de tradução ao usar estes diferentes tipos de luva, preparamos um plano experimental simples para medir a precisão com cada modelo. Em ambos os cenários de teste, a configuração de mão e a rotação são determinadas pela luva de dados, enquanto que o movimento da mão é independente delas e analisado apenas pelo Kinect.

Para avaliar a precisão da tradução de gesto para texto, criámos 3 *datasets* de teste diferentes, com 20, 50 e 100 palavras em cada. Cada uma destas palavras é um gesto de Língua Gestual Portuguesa. Os *datasets* são extensões uns dos outros, o de 100 palavras inclui as 50 palavras do *dataset* de 50, mais 50 novas, enquanto este último inclui as palavras do *dataset* de 20. Para os testes usamos sempre as mesmas 10 palavras.

Cada uma das palavras nestes *datasets* foi feita por um utilizador que sabe Língua Gestual Portuguesa. Cada palavra foi feita 5 vezes, e foi registado o número de acertos. O nosso tradutor mostra uma lista de palavras previstas ordenadas pela probabilidade, e consideramos a precisão baseada no resultado mais provável (descrito como *1st* na Tabela 1 e Tabela 2) juntamente com o segundo mais provável (descrito como *1st and 2nd* na Tabela 1 e Tabela 2). Fazemos o mesmo procedimento com as luvas 5DT e as IFG.

Os resultados na Tabela 1 e na Tabela 2 mostram que as luvas 5DT se portam significativamente melhor que as luvas IFG. Nos *datasets* de 20, 50 e 100 palavras, as 5DT atingiram uma precisão de 86%, 78% e 70%, respetivamente, quando considerados os dois resultados mais prováveis. As luvas IFG mostram resultados consideravelmente piores, atingindo 40%, 26% e 18%, respetivamente.

Apesar da margem entre os dois modelos parecer relevante, os resultados ainda são preliminares e ainda restringidos pelo problema anteriormente mencionado das luvas IFG usarem apenas giroscópios. Esperamos que, quando for resolvido esse problema e o resultado da previsão de configuração da mão for constante independentemente da rotação, tal como conseguimos com as 5DT, o resultado melhore significativamente. Estamos atualmente a trabalhar numa técnica baseada em gradientes para orientar os giroscópios pelo mesmo sensor.

VII. Melhoramentos Futuros

Agora que o sistema de tradução já funciona nas duas direções, estamos a definir melhoramentos futuros e correções para os próximos passos, para tentar passar de uma tradução básica para uma mais fluída.

Além de continuar a adicionar mais e mais palavras ao léxico das seis diferentes línguas gestuais, temos como objetivo melhorar o reconhecimento do contexto para tradução, primeiro na vertente texto para gesto, onde o processo de tradução já está mais avançado, e posteriormente no gesto para texto. Para isto é necessário continuarmos a trabalhar com intérpretes e utilizadores de língua gestual para tentar definir regras gramaticais e de sintaxe que possam ser aplicadas para melhorar a precisão da tradução. Para resolver

Tabela I

Precisão das luvas 5DT

Palavra de teste	20-sign dataset		50-sign dataset		100-sign dataset	
	1st only	1st and 2nd	1st only	1st and 2nd	1st only	1st and 2nd
Bom dia	5	5	4	5	4	5
Obrigado	5	5	5	5	5	5
Pós-Laboral	4	4	4	5	3	5
@	5	5	4	5	3	4
Avô	4	5	5	5	3	5
Avó	4	5	3	4	1	2
Secundário	0	0	0	0	0	0
Já	3	4	3	3	4	4
Janeiro	4	5	1	2	0	1
Dia	5	5	5	5	3	4
Precisão	78%	86%	68%	78%	52%	70%

Tabela II

Precisão das luvas IFG

Palavra de teste	20-sign dataset		50-sign dataset		100-sign dataset	
	1st only	1st and 2nd	1st only	1st and 2nd	1st only	1st and 2nd
Bom dia	2	2	0	0	0	1
Obrigado	3	4	1	2	0	0
Pós-Laboral	2	4	1	1	1	1
@	1	2	2	2	1	2
Avô	0	0	0	0	0	0
Avó	2	2	2	2	1	1
Secundário	0	0	0	0	0	0
Já	0	1	0	0	0	0
Janeiro	1	1	1	2	0	1
Dia	3	4	4	4	3	3
Precisão	28%	40%	22%	26%	12%	18%

este problema, uma das soluções que estamos a estudar é desenvolver um configurador de regras gramaticais que permita não só adicionar regras manualmente, mas também inferir automaticamente regras com base nas que foram registadas anteriormente.

No VSSO, um dos nossos principais objetivos neste momento é desenvolver mais o nosso avatar 3D para melhorar a sua apresentação, precisão dos movimentos e torna-lo mais simples de ler pelos utilizadores de língua gestual. Outra das nossas prioridades é alterar a forma como processamos as expressões faciais do avatar, e estamos já em reuniões com especialistas da área.

No que respeita à tradução de gesto para texto, os melhoramentos principais serão na classificação da configuração de mão, e estamos a testar continuamente novos classificadores e a estudar a luva IFG que ainda é novidade para nós, bem como implementar um sistema de probabilidades associadas ao grafo de tradução. Quando isto estiver implementado, o próximo passo é introduzir regras gramaticais e tentar reduzir ou simplificar o uso de dispositivos externos para a tradução.

VIII. Discussão

A acessibilidade e melhorar a experiência das pessoas com deficiência auditiva é objectivo principal do projecto ACE e do VirtualSign, não só na sala de aula mas em qualquer lugar em que estas pessoas não tenham possibilidade de ser acompanhadas poe um intérprete de língua gestual, e este produto pode ser uma grande ajuda nesse campo, especialmente se o processo de captura de movimentos e a tecnologia continuarem a melhorar. Outra das nossas prioridades principais é reduzir as barreiras de comunicação entre estudantes e professores, e é a razão pela qual vamos continuar a desenvolver mais ferramentas como o nosso *add-on* do PowerPoint.

Neste momento temos um léxico de 2,400 palavras em Língua Gestual Portuguesa, 460 para Língua Gestual Britânica, 550 para Língua Gestual Cipriota e Língua Gestual Eslovena, 600 para Língua Gestual Alemã, e 1200 para Língua Gestual Grega. Estes números vão aumentando a cada dia com os nossos parceiros, utilizadores de língua gestual e intérpretes, desses países a usar a ferramenta VSSO para introduzir mais palavras, o que nos permite traduções melhores em ambos os sentidos.

Testar a aplicação com dois modelos de luvas de dados diferentes tem sido muito útil para perceber os prós e contras de usar diferentes tipos de sensores na tradução de gesto para texto. Esperamos que os resultados melhorem quando for implementado o novo processo de obtenção de dados com as luvas IFG usando os giroscópios. O novo modelo de luva vai permitir reduzir os custos e melhorar a deteção da rotação da mão e do movimento, algo que pode vir a tornar redundante o uso do Kinect.

Agradecimentos

Este trabalho foi cofinanciado pela FCT – Fundação para a Ciência e a Tecnologia, no âmbito do projeto COMPETE2020 PTDC/IVC-COM/5869/2014, POCI-01-0145-FEDER-016584, pelo Instituto Superior de Engenharia do Porto e do GILT – Games, Interaction and Learning Technologies.

References

[1] J. Fellinger, D. Holzinger, and R. Pollard, "Mental health of deaf people," *Lancet*, vol. 379, no. 9820, pp. 1037–1044, 2012.

[2] Word Federation of the Deaf, "Our Work," 2019. [Online]. Available: http://wfdeaf.org/our-work/. [Accessed: 13-Sep-2019].

[3] T. M. M. de M. Martins, "A letra e o gesto: estruturas linguísticas em Língua Gestual Portuguesa e Língua Portuguesa," 2011.

[4] J. TE Richardson, L. Barnes, and J. Fleming, "Approaches to studying and perceptions of academic quality in deaf and hearing students in higher education," *Deaf. Educ. Int.*, vol. 6, no. 2, pp. 100–122, 2004.

[5] S. Riddell and E. Weedon, "Disabled students in higher education: Discourses of disability and the negotiation of identity," *Int. J. Educ. Res.*, vol. 63, pp. 38–46, 2013.

[6] and M. R. M. Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorrim, Tessa Verhoef, Christian Vogler, "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective . ACM.," *Proc. 21st Int. ACM SIGACCESS Conf. Comput. Access.*, 2019.

[7] S. Camargo Daroque, "Alunos surdos no Ensino Superior: uma discussão necessária," *Comunicações*, no. 2, pp. 23–32, 2011.

[8] R. Kheir and T. Way, "Inclusion of deaf students in computer science classes using real-time speech transcription," *ITiCSE 2007 12th*

Annu. Conf. Innov. Technol. Comput. Sci. Educ. - Incl. Educ. Comput. Sci., pp. 261–265, 2007.

[9] T. Oliveira, P. Escudeiro, N. Escudeiro, E. Rocha, and F. M. Barbosa, "Automatic sign language translation to improve communication," *IEEE Glob. Eng. Educ. Conf. EDUCON*, vol. April-2019, no. June, pp. 937–942, 2019.

[10] P. Escudeiro *et al.*, "Virtual Sign – A Real Time Bidirectional Translator of Portuguese Sign Language," *Procedia Comput. Sci.*, vol. 67, no. Dsai, pp. 252–262, 2015.

[11] J. Ulisses, T. Oliveira, P. M. Escudeiro, N. Escudeiro, and F. M. Barbosa, "ACE assisted communication for education: Architecture to support blind & deaf communication," in *IEEE Global Engineering Education Conference, EDUCON*, 2018, vol. 2018-April, no. May, pp. 1015–1023.

[12] Thomson Reuters, "Thomson Reuters Web of Science," *Web of Science*, 2018. [Online]. Available: http://apps.webofknowledge.com.

[13] M. Taskiran, M. Killioglu, and N. Kahraman, "A Real-Time System for Recognition of American Sign Language by using Deep Learning," 2018 41st Int. Conf. Telecommun. Signal Process. TSP 2018, pp. 1–5, 2018.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1725–1732, 2014.

[15] E. K. Kumar et al., "Training CNNs for 3-D Sign Language Recognition with Color Texture Coded Joint Angular Displacement Maps," IEEE Signal Process. Lett., vol. 25, no. 5, pp. 645–649, 2018.

[16] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks," IEEE Access, vol. 7, pp. 38044–38054, 2019.

[17] T. Yuan et al., "Large Scale Sign Language Interpretation," 2019 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2019), pp. 1–5, 2019.

[18] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based Sign Language Recognition without Temporal Segmentation," 2018.

[19] D. A. K. G. Anantha Rao, P.V.V. Kishore, A.S.C.S. Sastry and and E. K. Kumar, Selfie Continuous Sign Language Recognition with Neural Network Classifier, vol. 434, no. September. Singapore: Springer Singapore, 2018.

[20] Cao Dong, M. C. Leu, and Z. Yin, "American Sign Language alphabet recognition using Microsoft Kinect," in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 44–52.

[21] Z. Zhang, "Microsoft kinect sensor and its effect," IEEE Multimed., vol. 19, no. 2, pp. 4–10, 2012.

[22] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. 2005.

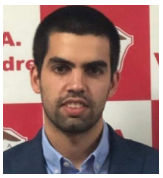
[23] P. Escudeiro et al., "Virtual sign translator," Proc. Int. Conf. Comput. Networks Commun. Eng. (ICCNCE 2013), no. Iccnce, pp. 290–292, 2013.

[24] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to arabic sign language recognition," IEEE Trans. Human-Machine Syst., vol. 44, no. 4, pp. 551–557, 2014.

Tiago André Queirós Oliveira completou a licenciatura em Engenharia Informática no Instituto Superior de Engenharia do Porto (ISEP) em 2014, e obteve o grau de mestre em Sistemas de Informação e Conhecimento em 2016. É estudante de doutoramento na Universidad Complutense de Madrid e está atualmente a trabalhar como investigador junior no laboratório Games, Interaction & Learning Technologies (GILT) no ISEP.

Nuno Filipe Fonseca Vasconcelos Escudeiro completou a licenciatura na Universidade Técnica de Lisboa, em Engenharia Informática e Electrónica, em 1991. Em 2004 obteve o grau de mestre em Sistemas de Suporte à Decisão e Análise de Dados, e em 2012 terminou o doutoramento em Informática na Universidade do Porto. Nuno é professor no Politécnico do Porto. Os seus interesses de pesquisa incluem *machine learning*, *text mining*. Nuno trabalhou na indústria de software entre 1989 e 2002, antes de ingressar definitivamente na via académica. Nuno é orientador do EO4GEO *sector skills alliance*, orientador da INNOTECS, a Associação Europeia de Escolas Técnicas, Vice-Predisente do Career Guidance, Vice-Presidente da Associação Europeira de Coordenadores Erasmus.

Paula Maria de Sá Oliveira Escudeiro, doutorada em Ciências da Computação/Tecnologias de Sistemas e Informação em Educação/Sistemas Multimédia. É professora no IPP-ISEP com vasta experiência em supervisão de projetos e avaliação, acumulada ao longo dos ultimos 25 anos. É a directora do grupo de investigação GILT (Games, Interacion & Learning Technologies) e coordenadora da área de pesquisa de Tecnologias de Aprendizagem. Presidente do Conselho da Associação de Jogos Sérios. Directora e Fundadora do Laboratório Multimádia do Departamento de Engenharia Informática. Subdirectora do Departamento de Engenharia Informática, Directora do Instiruto de Sistemas de Informação para Desenvolvimento Técnico, Vice-Presidente do Conselho Pedagógico, Membro do Comité Nacional para Projetos Europeis e Avaliadora Externa de Projetos Europeus. Directora do Centro para Desenvolvimento de Produtos Multimádia para Instituto Nacional. Membro do Subcomité de Qualidade de Avaliação no Instituto Superior de Engenharia do Porto. Directora do Programa de Pós-Graduação em Tecnologias de Informação e Comunicação. Tem vasta experiência como coordenadora/membro de mais de 20 projetos internacionais. Autora da patente nº 20091000045179, código 0198, processo 104557 K (Software Model Quality). Autorta da marca nº 20151000046313 código 059 (Virtual Sign). Autora da patente nº 20151000065572, processo 0198 (Bidirectional Translator Sign Language). Autora do modelo quantitativo de avaliação de qualidade de software - QEF (Quantitative Evaluation Framework). Vencedora de alguns prémios de investigação, incluíndo o prámio de Inclusão e Alfabetização Digital, entre outros.



Emanuel Vales Rocha obteve o gray de meste em Engenharia Informática no Instituto Superior de Engenharia do Porto (ISEP) do Politécnico do Porto. Foi investigador no GILT de Julho de 2017 até Janeiro de 2019 e é neste momento programador na Medidata e investigador externo do GILT. Os seus interesses de pesquisa são jogos sérios e software inclusivo.

Fernando Pires Maciel Barbosa (M`1976, SM`1982) nasceu no Porto. Licenciou-se na Universidade do Porto, em Engenharia Electroténica, em 1971. No mesmo ano, juntou-se ao staff da Faculdade de Engenharia Electrotécnica da Universidade do Porto (FEUP), como professor assistente. Em Outubro de 1976 juntou-se ao Departamento de Engenharia Electrotécnica na UMIST (Manchester). Em 1977 completou o mestrado e em 1979 o doutoramento, em Análise de Sistemas Energéticos. Atualmente é professor reformado na FEUP na área de Sistemas Energéticos. Os seus interesses de pesquisa são fiabilidade de sistemas energéticos, análise de sistemas energéticos a software relacionado. É autor de vários artigos e foi responsável por vários projetos na área de sistemas energéticos.