LibrasDetec: um Componente de Detecção de Movimentos para Karaokê em Libras

Manuella Aschoff Lima
Universidade Federal da Paraíba
João Pessoa, Paraíba
manuella.lima@lavid.ufpb.br

Daniel de França
Universidade Federal da Paraíba
João Pessoa, Paraíba
daniel.franca@lavid.ufpb.br

Arnor da Silva Neto
Universidade Federal da Paraíba
João Pessoa, Paraíba
arnor.neto@lavid.ufpb.br

Daniel Faustino de Souza
University Federal of Paraíba
João Pessoa, Paraíba
daniel@lavid.ufpb.br

Derzu Omaia
Universidade Federal da Paraíba
João Pessoa, Paraíba
derzu@lavid.ufpb.br

Tiago Maritan de Araújo
Universidade Federal da Paraíba
João Pessoa, Paraíba
tiago.maritan@lavid.ufpb.br

ABSTRACT

The entertainment market has been expanding rapidly, driven in part by increased internet access, which has also contributed to the growth of the digital games industry. However, despite this expansion, investments in accessibility remain limited, compromising the quality of games for people with disabilities. In the context of entertainment or serious games, particularly regarding accessibility for deaf users, motion detection emerges as a promising strategy to enable interaction through gestures in their native language. In this scenario, we propose LibrasDetec, a component for detecting and evaluating Brazilian Sign Language (Libras) gestures, integrated into the educational game Libraskê. The solution captures user gestures via webcam and compares them with reference videos, providing real-time scoring. To validate the proposed component, we conducted tests with 12 participants from three different user profiles: deaf individuals, interpreters, and non-signers. The results showed a satisfactory alignment between automated and human evaluations, especially among deaf users, suggesting that LibrasDetec is a viable approach to enhance accessibility and engagement in serious games focused on Libras.

KEYWORDS

serious game, motion detection, accessible game, sign language

1 INTRODUÇÃO

O ser humano é um ser social e, portanto, possui uma necessidade natural de conviver em comunidade, construir relações e se comunicar. No entanto, pessoas surdas nem sempre conseguem se comunicar na sociedade pois fazem uso de línguas de sinais — sistemas viso-espaciais com estruturas próprias e variações regionais e culturais.

No Brasil, a Língua Brasileira de Sinais (Libras) é a língua oficial da comunidade surda, reconhecida pela Lei nº. 10.436/2002[4], porém sua disseminação ainda enfrenta desafios históricos e sociais apesar de ser um instrumento fundamental de inclusão e expressão identitária[21].

Outras leis foram criadas visando reduzir as barreiras para o acesso de conteúdo para surdos, o que impulsionou vários avanços na área da acessibilidade digital, tais como desenvolvimento de softwares como VLibras [1] e Hand Talk[2], que atuam como intérpretes virtuais de conteúdo digital. No entanto, ainda são poucos os investimentos em produtos de entretenimento voltados para a comunidade surda, tais como em jogos e plataformas de streaming, sendo um exemplo o Libraskê.

O Libraskê é um jogo sério inspirado no karaokê, que tem como principal objetivo promover o contato lúdico e acessível com a Libras, cuja proposta visa incentivar a disseminação da Libras, tanto entre surdos quanto entre ouvintes, por meio de uma experiência interativa em que o protagonismo da língua de sinais é central. Sendo assim, o Libraskê permite que usuários surdos interajam utilizando sua língua nativa e que usuários ouvintes aprendam sua estrutura e expressividade em um ambiente visual e dinâmico, ressignificando também a forma como a musicalidade é vivenciada pela comunidade surda, por meio de ritmo, movimento e visualidade.

No entanto, visando viabilizar essa interação ativa em Libras, é necessário desenvolver uma solução tecnológica capaz de reconhecer os gestos dos usuários em tempo real e compará-los com os sinais de referência esperados pelo jogo. A partir desta motivação surge o LibrasDetec, um componente que utiliza visão computacional e inteligência artificial para identificar os movimentos corporais e manuais dos jogadores a partir da captura por webcam, fornecendo uma métrica de similaridade com base nos vídeos de referência.

A proposta do LibrasDetec busca superar as barreiras comuns no desenvolvimento de sistemas acessíveis para surdos, como o custo elevado de hardware específico e a baixa integração entre ferramentas de reconhecimento de gestos e jogos. Por meio da integração com o Libraskê, o LibrasDetec permite que os gestos sejam utilizados como comandos ativos e avaliáveis dentro do jogo, viabilizando uma experiência inclusiva e interativa.

Sendo assim, o objetivo geral deste trabalho é desenvolver e integrar ao jogo sério Libraskê um componente de processamento de imagens (LibrasDetec) que tem como função principal detectar movimentos de um usuário e fornecer uma métrica de avaliação com base na similaridade entre a interpretação dos movimentos em

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

[1] https://www.gov.br/governodigital/pt-br/vlibras
[2] https://www.handtalk.me/br/

Libras realizados pelo usuário e a interpretação de referência da música apresentada no jogo.

Este artigo encontra-se organizado em sete seções. Na Seção 2 é introduzido o embasamento teórico sobre o problema e solução apresentados; a Seção 3 apresenta os trabalhos relacionados; a Seção 4 descreve a solução proposta; a Seção 5 detalha a metodologia aplicada no desenvolvimento e validação; a Seção 6 traz os principais resultados obtidos e apresenta uma discussão sobre esses resultados. Por fim, são inseridas as considerações finais e questionamentos suscitados a partir desta pesquisa na Seção 7.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados conceitos fundamentais para o embasamento teórico da solução proposta.

2.1 Libraskê

O Libraskê é um jogo desenvolvido com o framework Unity 3D que apresenta uma proposta similar aos jogos de karaokê (ver Figura 1. Este jogo tem como destaque o foco na acessibilidade para pessoas surdas e o incentivo à aprendizagem de Libras de uma forma lúdica e engajadora. No Libraskê, os usuários interagem interpretando músicas que são apresentadas em Libras.

Os sinais são apresentados em tela por um avatar 3D, que auxilia na navegação e fornece uma referência de interpretação das músicas que são tocadas durante o jogo. Durante o *gameplay* o usuário, simultaneamente, é capaz de ver o vídeo capturado de sua câmera, para visualizar a comparação de movimentos com o avatar e obter um *feedback* visual do seu desempenho (ver Figura 1(f)).

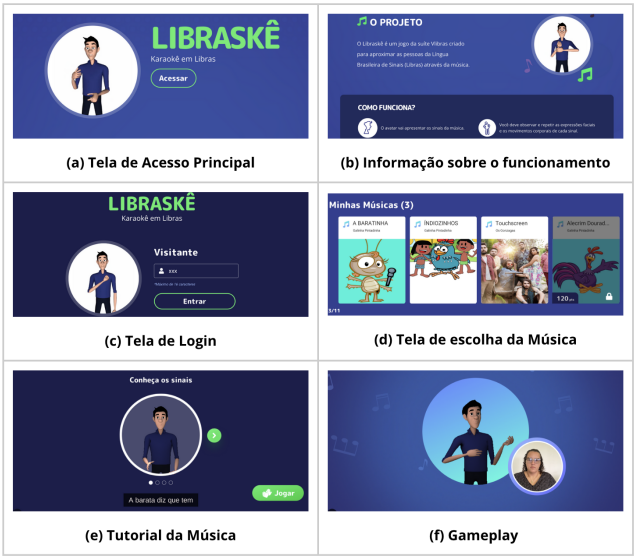

Figure 1: Telas de Acesso e Jogabilidade Libraskê

Como forma de manter o engajamento dos jogadores, o jogo possui um sistema de pontuação que permite que o usuário acumule uma moeda virtual capaz de efetuar ações no jogo, como desbloquear novas músicas (ver Figura 1-d) ou personalizar o avatar (ver Figura 2). No entanto, o sistema de pontuação não leva em consideração o desempenho do jogador na execução da interpretação.

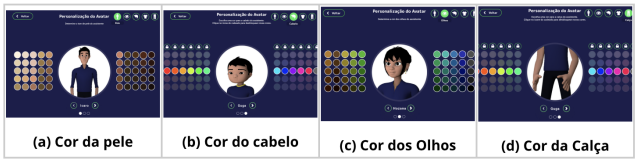

Figure 2: Telas de Personalização do Libraskê

Diante do exposto, e considerando que o Libraskê possui uma proposta de sistema de engajamento, o presente trabalho propõe a elaboração do componente de captura e processamento de imagens, o qual extrai dados do vídeo da webcam do usuário, os processa e retorna uma pontuação, em tempo satisfatório, baseada no desempenho do jogador na execução da sinalização em Libras. Tal proposta visa melhorar o engajamento, tornar o jogo mais lúdico e com maior potencial de incentivo ao entretenimento e aprendizado, posto que propõe uma pontuação baseada na corretude dos sinais executados.

2.2 MediaPipe

O MediaPipe[20] é um framework de código aberto desenhado para a construção de pipelines de inferência a partir de dados sensoriais do ambiente. A biblioteca fornece soluções pré-treinadas baseadas em Aprendizagem de Máquina (AM) para identificação de rosto, pose e mãos humanas a partir de imagens. Além disso, o MediaPipe, tem como característica poder ser executado em dispositivos móveis o que viabiliza o seu uso na aplicação proposta, posto, que fornece respostas satisfatórias em tempo real e com pouca exigência de hardware para processamento.

Considerando a solução proposta, além da captura de movimentos do usuário, outra funcionalidade fundamental para o componente desenvolvido é a capacidade de determinar a qualidade dos movimentos realizados pelo usuário. Para a realização da comparação entre as poses de referência e as poses dos usuários, foi estudado o algoritmo *Dynamic Time Warping*, apresentado na sessão seguinte.

2.3 Dynamic Time Warping

O *Dynamic Time Warping (DTW)* é um algoritmo capaz de medir a similaridade entre duas sequências temporais[23]. No DTW, as sequências são comparadas de forma não-linear em relação ao tempo, com o objetivo de minimizar a distância euclidiana entre as séries após o alinhamento ideal, compensando por certas variações não-lineares na dimensão do tempo. Sendo assim, usando este algoritmo, é possível observar a ocorrência ou determinar a similaridade de um mesmo padrão presente em duas séries, ainda que em cada uma delas o padrão tenha sido construído em diferentes períodos ou velocidades.

A seleção do DTW, se deu por este ser um algoritmo popular e que tem aplicações viáveis em diversas áreas e por ser implementado com programação dinâmica com algoritmo otimizado e com baixo consumo de recursos.

3 TRABALHOS RELACIONADOS

Nesta seção são apresentados trabalhos relacionados à pesquisa e desenvolvimento da solução proposta, destacados em trabalhos correlatos à detecção de gestos e reconhecimento de língua de sinais, o processo de comparação de sequências e cálculo de similaridade de gestos, o uso de abordagens baseadas em jogos sérios e a sua aplicação no processo de ensino de línguas de sinais.

Pesquisas recentes com abordagens baseadas em Visão Computacional e Ferramentas de Detecção de Poses têm impulsionado o Reconhecimento de Língua de Sinais (RLS) baseado em visão, a exemplo da introdução de um dataset multi-view de grande escala demonstrando ganhos na acurácia em comparação com dados de visão única, ressaltando a importância de múltiplas perspectivas na captação de nuances dos sinais[12]. Além disso, modelos baseados em Transformers, como VideoMAE e TimeSformer têm se mostrado promissores no processo de RLS, representando um avanço relevante na modelagem da natureza dinâmica da língua de sinais[5].

A detecção precisa de pontos-chave (landmarks) do corpo, mãos e rosto é fundamental para o reconhecimento e a avaliação da língua de sinais e avanço da visão computacional, especialmente com bibliotecas de código aberto como MediaPipe, tem permitindo o desenvolvimento de sistemas mais acessíveis e confortáveis. Um exemplo do uso de MediaPipe para reconhecimento de língua de sinais demonstraram a eficácia desta para a coleta de dados de pose, utilizando-a em conjunto com algoritmos de aprendizado de máquina, como Support Vector Machine (SVM), para classificar gestos de diversas variantes de língua de sinais, incluindo americana, indiana, italiana e turca [16]. Esta proposta alcançou uma acurácia superior a 99%, para o reconhecimento de letras do alfabeto e números o que aponta a relevância deste estudo para o LibrasDetec é a validação do MediaPipe como uma ferramenta robusta para extrair coordenadas de pose sem a necessidade de sensores vestíveis, o que contribui para uma solução mais confortável e econômica.

Outro aspecto importante na solução LibrasDetec são os algoritmos de comparação de sequências e avaliação de movimentos. Como exemplo, encontra-se o DTW sendo aplicado na verificação de assinaturas off-line apresentando eficácia na identificação de similaridade entre vetores de características, mesmo na presença de variações temporais na execução das assinaturas[24]. Além disso, modificações no algoritmo DTW foram propostas para melhorar sua estabilidade, alcançando uma taxa de erro de aproximadamente 2% com o DTW modificado, em contraste com 29% para o DTW original[24].

A evolução do DTW tem levado à sua integração com arquiteturas de aprendizado profundo através do uso de kernels DTW dentro de redes neurais para extração de características [26]. Essa abordagem permite que o DTW seja otimizado e integrado em pipelines de aprendizado profundo. Outra abordagem combina a rede YOLOv5 para detecção de mãos com o algoritmo DTW para reconhecimento de gestos em mídias digitais [19] e outra comparou o DTW com outras técnicas para busca em léxicos de língua de sinais, mostrando que o DTW pode prever um sinal correspondente com boa acurácia mesmo com variações na execução do sinal [15].

Diante disso, percebe-se que o DTW é robusto a variações de velocidade e duração de execução de sequências sendo uma ferramenta poderosa para comparar sequências de dados convertidas em estruturas vetorizadas, atestando sua aplicação no contexto deste trabalho.

Por fim, a a gamificação e acessibilidade na aprendizagem de Língua de Sinais (LS) tem se mostrado uma estratégia eficaz para aumentar o engajamento e a motivação na aprendizagem de novas habilidades. A aplicação de elementos de jogo em contextos educacionais pode tornar o processo de aquisição de uma nova língua mais interativo e divertido. No campo de jogos voltados ao ensino de LS, observa-se uma diversidade de abordagens que aliam interatividade, tecnologia e engajamento lúdico para promover aquisição linguística e inclusão.

O SIGNIFY, é um exemplo desta tecnologia, pois utiliza reconhecimento gestual assistido por aprendizado de máquina para ensinar Língua de Sinais Italiana (LIS) a crianças por meio de um serious game, validado com bons índices de usabilidade, engajamento e eficácia pedagógica [25]. Similarmente, verifica-se o uso do role play games (RPG) para o ensino de American Sign Language (ASL) a estudantes do ensino fundamental, com resultados de 50 a 100 palavras aprendidas e ganhos em motivação, cooperação social e disposição para continuar o aprendizado [10]. Assim como, a combinação de dinâmicas de jogo com feedback visual e prática gestual ativa em LIBRAS, demonstrando viabilidade técnica e aceitação pelos educadores [2].

A proposta precursora no desenvolvimento de jogos educativos para ASL foi um jogo para crianças surdas que tinham ASL como língua principal, onde as crianças imitavam sinais apresentados em vídeo para interagir com um mascote animado[17]. Contudo, a avaliação do reconhecimento de sinais foi inicialmente simulada usando o método "Mágico de Oz", no qual um intérprete humano oculto fornecia as respostas do sistema. Este método revelou um desafio crítico: a subjetividade e a inconsistência da avaliação humana, com diferentes avaliadores apresentando critérios distintos, dando um indicativo que uma avaliação automatizada com possibilidade de feedback em tempo real possa ser uma alternativa mais eficaz. Este ponto é crucial para o LibrasDetec, que busca uma avaliação automatizada proporcionando feedback em tempo real.

Alinhado ao objetivo de integrar um componente de avaliação ao jogo sério Libraskê, o LibrasDetec representa um avanço ao unir técnicas de detecção de pose por visão computacional (MediaPipe), comparação de sequências temporais (DTW) e geração automatizada de referências. Essa integração ocorre em um ambiente gamificado de karaokê em Libras, que alia entretenimento e aprendizagem de forma acessível, proporcionando uma experiência imersiva e inclusiva para usuários surdos e ouvintes interessados em aprender a língua.

4 SOLUÇÃO PROPOSTA

O LibrasDetec tem como objetivo promover a avaliação, com base no desempenho do jogador, da interpretação em Libras das músicas no jogo Libraskê. Sendo assim, avalia o usuário a partir da replicação dos sinais (ou movimentos) exibidos na tela no jogo (ver Figura 1(d)). A partir desses movimentos, o sistema captura as imagens do usuário e as processa realizando uma comparação com a referência.

Após a comparação do vídeo do jogador com o vídeo de referência o sistema retorna ao usuário uma pontuação, a qual é calculada de acordo com a semelhança entre a pose capturada do usuário e a pose de referência previamente contida no sistema.

Para fornecer uma métrica de qualidade ao usuário em tempo de processamento adequado, o sistema requer que os dados de referência sejam previamente obtidos e armazenados na aplicação ou em um servidor em nuvem. Em seguida, é preciso criar um serviço para processar e armazenar as coordenadas de pose de um vídeo de referência em um formato que possa ser posteriormente usado para a avaliação, chamado de gerador de referências. Depois, é necessário implementar o componente que captura e extrai as coordenadas de pose do vídeo do usuário e as processa por meio de um algoritmo que compara e calcula a métrica de similaridade entre os dados de pose capturados e os dados previamente armazenados no sistema, chamado de comparador de coordenadas.

4.1 Gerador de Referências

O Gerador de Referências foi implementado em Python, cujo serviço funciona a partir do upload de um vídeo e tem como função extrair as coordenadas de pose dos frames de vídeo com uso da biblioteca Mediapipe e, em seguida, armazena os dados extraídos em um objeto no formato JSON para uso posterior.

Adicionalmente, o serviço também realiza marcação nos frames do vídeo, ou seja, desenha indicadores visuais das coordenadas identificadas nas imagens por meio de pontos e os liga com linhas. A partir destas imagens marcadas, é possível gerar um novo vídeo com uma silhueta similar à do esqueleto humano representando os movimentos identificados ao longo do vídeo (ver Figura 3).

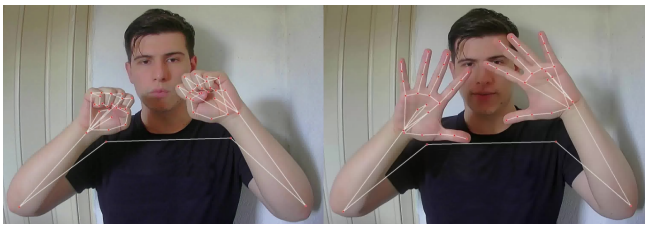

Figure 3: Marcação de coordenadas em frames de vídeo

O gerador de referências possui fluxo funcionando de forma independente (ver Figura 4), separado do comparador de coordenadas. Sendo assim, o processo de geração de coordenadas deve ser executado para todos os vídeos que serão usados como referência e os artefatos gerados (conjunto de dados e vídeo demarcado) precisam ser transferidos manualmente para o ambiente da aplicação, que fará uso do comparador de coordenadas.

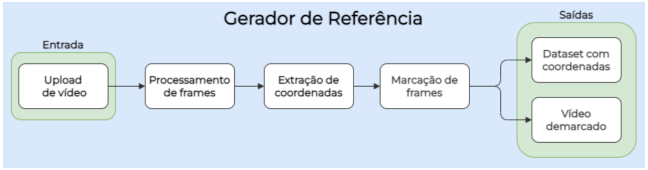

Figure 4: Fluxo de dados do gerador de referência

4.2 Comparador de Coordenadas

O comparador de coordenadas é um serviço que se baseia na leitura de frames obtidos a partir de vídeo da webcam do usuário e foi implementado usando JavaScript com auxílio do framework React e a biblioteca Mediapipe com suporte às funções nativas de entrada de imagens, sendo implementada apenas uma função para ser executada sempre que um novo frame do vídeo for processado.

Como o Mediapipe atinge processamento de cerca de 30 FPS variando por dispositivo, o algoritmo de comparação implementado seria executado um alto número de vezes por segundo, gerando um grande volume de processamento de dados. Sendo assim, para evitar processamento excessivo e manter desempenho satisfatório, foi estipulado um intervalo de tempo entre as chamadas do algoritmo de comparação. Por padrão, um frame resultante é capturado e processado a cada 500 milissegundos, sendo esse intervalo um delimitador, tanto para o gerador de referências, como para o algoritmo comparador.

O comparador de coordenadas funciona por meio de sessões de execução, cada sessão correspondendo ao progresso em um determinado trecho de vídeo previamente processado. A cada intervalo, o algoritmo comparador é executado recebendo o frame que foi processado pelo Mediapipe e o conjunto de coordenadas de pose e mãos identificados na imagem. Em seguida, são obtidas as coordenadas do frame de referência e, após a obtenção dos dados, o algoritmo realiza a comparação dos conjuntos de coordenadas usando o algoritmo DTW (Ver Figura 5.)

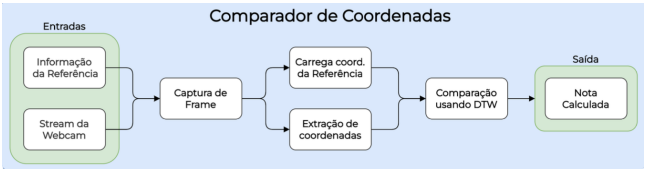

Figure 5: Fluxo de dados do comparador de coordenadas

Cada comparação resulta em um valor decimal no intervalo [0, 1], que representa a distância média registrada entre os dois conjuntos de coordenadas comparados pelo DTW. Por apresentar alta densidade de casas numéricas, é feita uma conversão de escala e o valor obtido é convertido para o intervalo [0, 10], com sensibilidade de duas casas decimais. Esta conversão aumenta a visibilidade das casas decimais mais impactantes na medida de desempenho e torna o valor resultante mais legível para apresentação ao usuário. Após cada cálculo, o resultado individual é enviado para o indicador visual de desempenho na interface do jogo e é armazenado para uso no fim da sessão. Ao fim da sessão, é calculado o valor médio das pontuações obtidas em todos os frames processados na sessão finalizada, que pode então ser apresentado para o usuário como pontuação final ou tratado pela aplicação que faz uso do componente.

5 METODOLOGIA

Nesta seção serão detalhados os métodos e estratégias adotadas para o desenvolvimento da solução proposta, bem como o processo de avaliação e coleta de resultados.

5.1 Seleção de Tecnologias

Dada a complexidade de processamento do componente, foi realizado um levantamento e comparativo entre algumas bibliotecas do estado da arte que utilizam métodos de AM para a captura de pose em imagens, tais como: OpenPose [6], AlphaPose [13] e Mediapipe [20].

No processo de avaliação das bibliotecas, foi usada uma amostra de vídeos em Libras para teste, que foi processada por todas as bibliotecas e, durante o processamento, foram avaliados fatores como: requisitos mínimos de hardware, tempo de processamento, precisão dos resultados e complexidade de integração.

Durante esta avaliação exploratória, foram observadas limitações no OpenPose e AlphaPose, que, devido à inicialização de modelo em tempo de execução, exigem, para processamento em tempo satisfatório, que os dispositivos sejam equipados com GPUs Nvidia, que suportem a plataforma de processamento paralelo CUDA. Já o MediaPipe suporta renderização em frameworks nativos como OpenGL e apresentou desempenho satisfatório com requisitos de hardware e tempos de resposta baixos, alcançando taxas de processamento acima de 10 FPS no hardware de menor desempenho usado nos testes. Diante disso, optou-se pelo Mediapipe para o desenvolvimento da solução proposta.

5.2 Integração com o Libraskê

O Libraskê é uma aplicação complexa que funciona unindo vários serviços, sendo seu componente principal um jogo desenvolvido com o framework Unity3D que é executado em um endereço web via OpenGL. Como o Mediapipe também deve ser executado em navegador via OpenGL e todo o fluxo de dados de uma sessão de jogo fica contida apenas dentro do ambiente de execução Unity, implementado na linguagem C. Adotou-se uma estratégia de integração remota entre o ambiente de jogo e o componente comparador de coordenadas.

Para tal, o componente de pontuação foi instanciado em um servidor REST e as imagens do vídeo do usuário passaram a ser capturadas pela aplicação Unity e enviadas para o servidor através de chamadas HTTP. A aplicação envia para o servidor um comando de início de sessão de jogo, que inicializa o armazenamento das pontuações obtidas nas avaliações dos frames, que são enviados em seguida. Ao fim da sessão de jogo, analogamente, é enviado para o servidor o comando de encerramento de sessão, que retorna à aplicação a pontuação média obtida pelo usuário (Ver Figura 6).

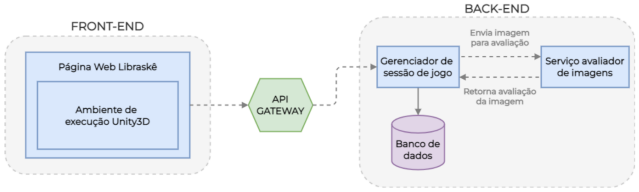

Figure 6: Integração entre Libraskê e LibrasDetec

Este tipo de abordagem permite que eventuais alterações e manutenções no serviço de avaliação sejam realizadas de forma independente à execução do jogo.

5.3 Processo de validação do componente

Para avaliar a solução proposta, realizou-se a comparação da similaridade entre as avaliações do sistema e as avaliações de um agente humano especialista, o intérprete de Libras, posto que o sistema é baseado em língua de sinais.

Durante o teste, o usuário deve logar no Libraskê (Figura 1(a), (b) e (c)), depois escolhe a música (Figura 1(d)), em seguida vê o tutorial que contém os principais sinais da música (Figura 1(e)), e por fim realiza a interpretação (ver Figura 1(f)).

O recrutamento dos participantes foi realizado via redes sociais, o teste foi aprovado por comitê de ética e todos os participantes assinaram termo de livre consentimento. Além disso, os participantes do teste foram recrutados sob três perfis: (1) Surdos; (2) Ouvintes - Intérpretes de Libras; e (3) Ouvintes - Leigos. O grupo de usuários surdos foi avaliado com ausência de áudio e, consequentemente, de ritmo das músicas. Os grupos de ouvintes avaliaram tendo acesso ao som, letra e ritmo da música. Por fim, foi selecionado um quarto grupo também composto por intérpretes de Libras, que avaliou as sessões de jogo gravadas com os grupos de teste e respondeu a formulário digital, fornecendo notas para as interpretações das músicas e indicando os fatores que contribuíram com a atribuição da nota.

6 RESULTADOS E DISCUSSÕES

A avaliação consistiu de três fases, sendo a primeira composta pelas sessões de playtest, realizadas em ambiente controlado e com a presença de um analista de qualidade, um intérprete e doze usuários. A segunda fase do experimento, também realizada em ambiente controlado, foi aplicada por um analista de qualidade, junto aos três intérpretes avaliadores, os quais observaram, cada um deles, os 36 vídeos e forneceram uma pontuação para cada interpretação da música, com base na correta interpretação e execução dos sinais. Sendo assim, foram realizados testes com 12 jogadores, cada um executando 3 músicas, que foram avaliadas por 3 intérpretes, totalizando 108 avaliações humanas do desempenho de usuários dentro do jogo. Por fim, a terceira fase consistiu na realização de análises estatísticas visando examinar a correspondência entre as notas do LibrasDetec e dos avaliadores humanos.

6.1 Observações gerais sobre as sessões de jogo

Durante a aplicação dos testes, os analistas de qualidade realizaram observações e coletaram feedback espontâneo sobre a interação com o jogo e que apontaram alguns comportamentos relevantes dos usuários, tais como: (1) Alguns usuários solicitaram repetição da sessão de jogo, apenas com o intuito de melhorar a pontuação, mesmo essa não sendo considerada na pesquisa; e, (2) Os jogadores intérprete e os avaliadores dos vídeos apontaram a falta de sincronia entre a música e os sinais do vídeo de referência em algumas das músicas do teste.

Essas impressões apontam para uma boa capacidade do Libraskê em despertar o engajamento do jogador, o que também pode indicar que o jogo tem potencial educacional, visto que os usuários que repetiram suas sessões apresentaram desempenho superior nas novas interações.

Considerando o relato sobre a falta de sincronia, alguns usuários relataram dificuldade em manter a concentração nos sinais apresentados no vídeo, posto que a interpretação e a ordem gramatical dos sinais em Libras difere dos da Língua Portuguesa. Sendo assim, há um indicativo de que múltiplos estímulos geraram dificuldade quando a sequência de palavras da música que era ouvida (música tocada) não condizia com a sequência de sinais apresentados, fato este que é comum na Libras. Analogamente, os intérpretes obtiveram desempenho superior nas músicas onde se percebia uma maior literalidade da letra da música e que, consequentemente, levava à sincronia entre a letra da música e os sinais apresentados na interpretação.

Outra observação relevante identificada pela equipe de aplicadores dos testes e que auxiliou a endereçar uma questão de desempenho técnico foi que nos grupos que possuem fluência em Libras, a familiaridade com os movimentos da língua fez com que esses usuários realizassem movimentos mais rapidamente, causando o fenômeno de "borrão" nas imagens capturadas. Tal fato impactou, para estes jogadores, em notas mais punitivas endereçadas pela solução proposta, pois dificultou a captura das coordenadas de mão, levando a situações de "falso negativo" e, consequentemente, reduzindo de forma indevida a pontuação final destes usuários.

6.2 Resultados gerais e Análise da avaliação

As notas obtidas pela solução proposta e a média das notas fornecidas pelos avaliadores humanos, utilizando como métrica o módulo da distância entre estão apresentadas nas Tabelas 2 e 1 e os resultados são discutidos a seguir.

Table 1: Nota média do componente (NMC), Média das Notas dos Avaliadores (MNA) e Média da Distância entre as Notas (MD), por Grupo de Usuário

Grupo	NMC	MNA	MD
Intérpretes	7,16	7,36	0,81
Leigos	7,20	7,58	0,83
Surdos	7,22	7,47	0,61
Todos	7,19	7,47	0,75

A média das distâncias evidenciadas na Tabela 1, apresenta que a distância média entre todas as notas é de 0,75 e observa-se que a média geral do componente foi 7,19, enquanto que a média das notas dos avaliadores foi 7,47, evidenciando um maior rigor na nota da solução proposta. Além disso, ainda na Tabela 1 percebe-se que há um maior alinhamento entre as notas do componente e a média das notas dos avaliadores no grupo dos surdos e um maior desalinhamento no grupo dos leigos, sugerindo que o componente pode ter um alinhamento quanto ao perfil do usuário.

Na Figura 7 percebe-se que 24 usuários, ou seja, 67% apresentaram valor da distância menor que 1, enquanto 12 usuários, ou seja, 33% apresentaram distância maior ou igual a 1 e menor que 2. Dentre as distâncias obtidas o maior valor foi 1,51 e o menor 0,03. Observa-se ainda que dentre as distâncias maiores ou igual a 1, 25% ocorreram na música "Borboletinha", 33% na música "Maria Bonita", e 42% no "Hino Nacional", o que sugere que músicas mais

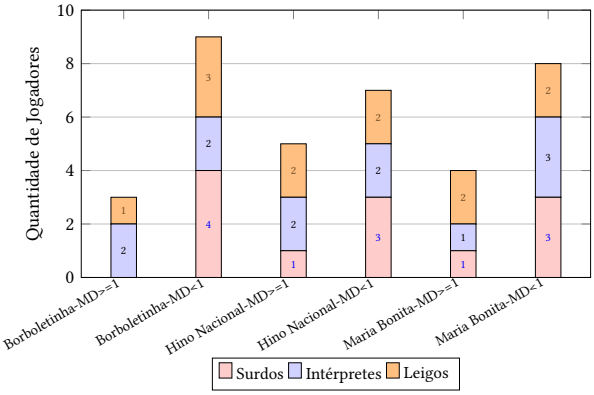

Figure 7: Distribuição de jogadores com Distância >=1 e Distância < 1 por grupo e música.

complexas e com maior quantidade de sinais há mais divergência entre a nota do componente e dos avaliadores.

Ainda considerando as distâncias maiores ou iguais a 1, percebe-se que 84% dos usuários foram ouvintes e apenas 16% foram surdos. Uma provável explicação para essas divergências, no caso dos intérpretes, é que a falta de sincronia entre os sinais e o áudio da música ocasionou a perda de concentração, como mencionado na seção anterior. Além disso, também identificou-se que, dentre o grupo de intérpretes, o comparador de coordenadas identificou "falsos negativos", conforme mencionado na seção anterior.

Table 2: Nota média do componente (NMC), Média das Notas dos Avaliadores (MNA) e Média da Distância entre as Notas (MD), separada por Música e por Grupo de Usuário

Música	Grupo	NMC	MNA	MD
Borboletinha	Surdos	7,46	7,84	0,45
	Intérpretes	7,29	8,13	0,84
	Leigos	6,89	6,99	0,65
	Geral	7,21	7,65	0,65
Hino Nacional	Surdos	6,96	7,69	0,73
	Intérpretes	6,19	5,32	0,96
	Leigos	7,09	8,06	0,97
	Geral	6,75	7,02	0,89
Maria Bonita	Surdos	7,25	6,89	0,63
	Intérpretes	8,02	8,63	0,62
	Leigos	7,61	7,69	0,87
	Geral	7,62	7,74	0,71

Outra observação é que, percebe-se dentre as distâncias maiores ou igual a 1, 66% das notas fornecidas pelo componente foram mais punitivas, ou seja, menores do que a nota fornecida pelos avaliadores. No tocante a todas as notas dos testes, identifica-se 69% das notas do componente foram mais punitivas o que indica que o componente tende a perceber divergências de forma mais precisa, mesmo que pequenas, entre os movimentos da sinalização e de referência punindo-as.

A Tabela 1 traz que, dentre os usuários que realizaram o teste, os surdos foram os que obtiveram menor média entre as distâncias das notas das músicas. Aprofundando esta análise, na Tabela 2, percebe-se que em duas músicas os surdos mantiveram a menor média entre as distâncias e apenas em uma música eles ficaram com o segundo melhor resultado, porém ficando maior apenas 0,01. Sendo assim, há um indicativo de que os surdos são mais fluentes nos sinais e não foram influenciados pelo seu ritmo, portanto, conseguiram sinalizar de forma mais precisa e síncrona ao vídeo de referência.

Observa-se também na Tabela 2 que a música com a maior média entre as distâncias foi o "Hino Nacional" o que pode ser explicado, pois os sinais são mais complexos, diversos e em maior quantidade, logo precisam ser executados de forma mais rápida para se adequar ao tempo da música e ao ritmo. Além disso, observa-se que a música Borboletinha apresentou a média entre as distâncias igual a 0,65, tendo a menor média entre os surdos e a maior entre os intérpretes. Já no "Hino Nacional", verifica-se que a média entre distâncias é igual a 0,89, porém desta vez o grupo que teve a maior média entre as distâncias foi o dos leigos. Por fim, na música "Maria Bonita" a média entre as distâncias foi de 0,71, sendo a menor a dos intérpretes e a maior a dos leigos. Tais informações nos apontam que as notas do componente e dos avaliadores tiveram menor divergência na música "Borboletinha", demonstrando maior alinhamento entre as notas calculadas pela solução proposta e dos intérpretes devido à simplicidade e caráter repetitivo dos sinais presentes nesta música.

Observa-se ainda, que a música Maria Bonita obteve alinhamento relativamente melhor entre as distâncias médias dos grupos, com a maior parte das ocorrências de distância entre os resultados sendo causadas por avaliações dos intérpretes superiores à avaliação do componente. Uma possível explicação é uma maior sincronia entre a música e o vídeo de referência, melhorando na percepção de ritmo, especialmente entre os ouvintes.

Visando entender os fatores que mais impactaram na avaliação dos intérpretes na segunda fase do experimento, o questionário de avaliação solicitava que apontasse os principais critérios que impactaram para a aferição de cada nota. A Figura 8 exibe a relação dos motivos assinalados e percebe-se que cerca de 89% das sessões avaliadas incluíram o critério de sincronia entre usuários e avatar como fator determinante para a nota. A qualidade da sinalização foi considerada em 59% das sessões, ritmo da interpretação em cerca de 30% e outros motivos (em sua maior parte indicando erros críticos na sinalização) foram apontados em cerca de 3% das sessões.

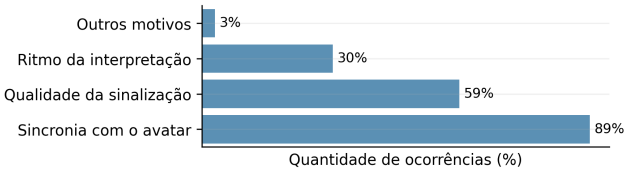

Figure 8: Principais critérios considerados na avaliação

O critério mais assinalado pelos intérpretes avaliadores demonstra o alinhamento de expectativas entre a avaliação automatizada e a avaliação humana, uma vez que a sincronia entre a sinalização do usuário e o vídeo de referência é o principal critério utilizado no cálculo das pontuações determinadas pelo componente implementado.

Diante do exposto, percebe-se maior alinhamento entre as notas do componente e dos avaliadores para o grupo de usuários surdos, especialmente na música Borboletinha. Já no grupo de usuários intérpretes, percebe-se mais dispersão das notas apontando para uma necessidade de melhoria técnica do componente para evitar a detecção de "falsos negativos" pelo comparador de coordenadas.

6.3 Relação entre o componente e a avaliação dos intérpretes.

Com o objetivo de examinar o grau de correspondência entre os valores produzidos pelo componente automatizado e as notas atribuídas pelos avaliadores humanos, foi utilizado um modelo de Regressão linear [3].

Na análise exploratória, a visualização do gráfico de dispersão (ver Figura 9(a)) sugeriu uma relação não linear entre as variáveis estudadas. Para investigar essa possibilidade, utilizou-se uma linha de suavização LOESS (*Locally Estimated Scatterplot Smoothing*) sobre o gráfico de dispersão, conforme proposto por [8]. Adicionalmente, aplicou-se o teste de especificação de Ramsey RESET (*Regression Equation Specification Error Test*), que avalia formalmente a presença de não linearidade na equação de regressão. Os resultados indicaram um valor de estatística $F = 7{,}39$ com p-valor de 0,010, o que é estatisticamente significativo ao nível de 5% ($p < 0{,}05$), reforçando a necessidade de aplicar transformações às variáveis para melhorar o ajuste do modelo [22].

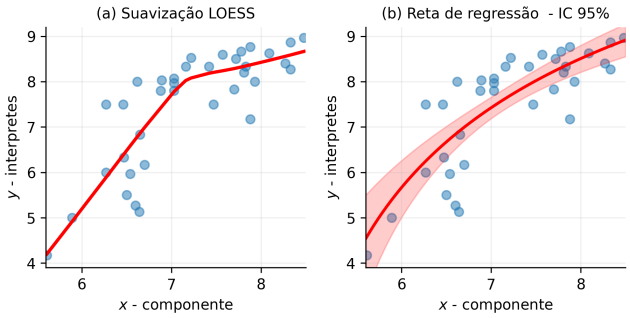

Figure 9: Gráficos de dispersão com suavização LOESS (a) e regressão com intervalos de confiança de 95% (b)

Com esses resultados, foram testadas diferentes transformações sobre as variáveis dependente e independente e a combinação que apresentou o melhor ajuste, conforme o critério de informação de Akaike (AIC), foi a inversa da variável independente ($1/x$) e o quadrado da variável dependente (y^2) [1, 3]. A aplicação dessas transformações resultou em coeficientes estatisticamente significativos ao nível de 1%, conforme demonstrado na Tabela 3. Para facilitar a interpretação visual do modelo ajustado, os valores previstos foram revertidos para as escalas originais e representados na Figura 9(b), onde se observa a linha de regressão estimada acompanhada de seus respectivos intervalos de confiança de 95%.

[3]Regressão linear simples que busca modelar a relação entre duas variáveis contínuas, uma dependente e uma independente, por meio de uma linha reta que minimiza a soma dos quadrados dos resíduos.

Table 3: Coeficientes estimados da regressão para a avaliação dos intérpretes (y) em função do valor do componente ($1/x$)

$\hat{\beta}$	Coeficiente	t-valor	p-valor
$\hat{\beta}_0$	193.2834	11.776	0.000
$\hat{\beta}_1$	-967.0889	-8.324	0.000

Sendo assim, o modelo ajustado apresentou bom desempenho explicativo, com um coeficiente de determinação ajustado (R^2_{ajustado}) de 0,661 indicando que aproximadamente 66,1% da variância da variável dependente transformada (y^2) é explicada pela variável independente transformada ($1/x$). A significância global do modelo é confirmada pela estatística $F = 69{,}29$, associada a um p-valor inferior a 0,01 ($p = 1{,}02 \times 10^{-9}$), permitindo rejeitar a hipótese nula de que todos os coeficientes são simultaneamente nulos [14].

Na avaliação dos pressupostos do modelo, os testes formais de normalidade, como Jarque-Bera ($p = 0{,}588$) e Omnibus ($p = 0{,}584$), além dos valores de assimetria (Skew = $-0{,}288$) e curtose (Kurtosis = 2,387), indicam que a hipótese de normalidade dos resíduos não pode ser rejeitada ao nível de significância de 5%. Complementarmente, a inspeção visual do Q-Q Plot (Figura 10(a)) corrobora a adequação da distribuição dos resíduos para fins inferenciais [7, 11, 18].

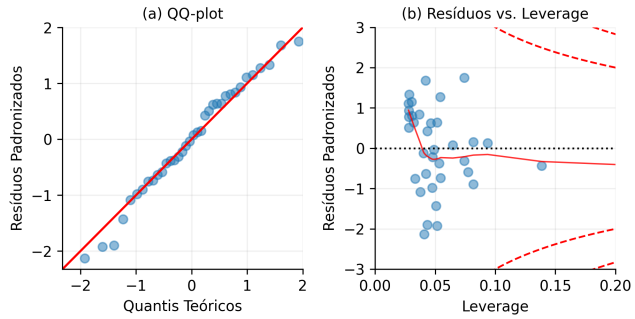

Figure 10: Gráficos de Diagnóstico para Validação dos Pressupostos do Modelo de Regressão

Além disso, a Figura 10(b) mostra que não há observações excessivamente influentes ou potenciais *outliers*. No gráfico de resíduos padronizados versus alavancagem, nenhum ponto ultrapassa os limites da Distância de Cook, representados pelas linhas tracejadas vermelhas [9], indicando que o ajuste do modelo não é sensivelmente influenciado por observações individuais.

Portanto, a modelagem estatística revelou uma relação significativa e sistemática entre os valores fornecidos pelo componente automatizado e as notas atribuídas pelos avaliadores humanos, que pôde ser adequadamente representada por meio de regressão. A não linearidade observada inicialmente entre as variáveis foi eficazmente tratada com transformações matemáticas, resultando em um modelo com bom poder explicativo ($R^2_{\text{ajustado}} = 66{,}1\%$). Em suma, os resultados fornecem evidência estatística robusta de que o componente automatizado reflete critérios compatíveis com as avaliações humanas, especialmente no que diz respeito à sincronia dos sinais, aspecto central na avaliação qualitativa da interpretação.

7 CONSIDERAÇÕES FINAIS

Este trabalho apresentou o LibrasDetec, um componente de software proposto para identificar movimentos e sinais manuais realizados por usuários em tempo real, com o objetivo de viabilizar experiências mais acessíveis e interativas em jogos do tipo karaokê baseados na Libras. A proposta se insere no contexto do desenvolvimento do jogo Libraskê, voltado à promoção da inclusão e da disseminação da Libras em contextos educacionais e de entretenimento.

Os resultados obtidos até o momento demonstram o potencial da abordagem baseada em visão computacional para detectar gestos da Libras em sincronia com conteúdos musicais e visuais. A aplicação do LibrasDetec foi avaliada de forma preliminar em testes com usuários, revelando indícios promissores de engajamento e de reconhecimento dos sinais, ainda que algumas limitações tenham sido observadas, especialmente relacionadas à variação individual na realização dos sinais e janela de captura dos frames.

Durante as sessões de jogo, os usuários apresentaram engajamento elevado e declararam a intenção de repetir suas sessões com o objetivo de obter pontuações melhores, evidenciando o propósito de desafio e recompensa que o sistema de pontuação tem a oferecer. Durante a avaliação das sessões por intérpretes, o principal critério considerado relevante para as avaliações foi o de sincronia entre usuário e avatar, indicando expectativas alinhadas com o principal critério de avaliação do software desenvolvido.

No entanto, também foram detectadas algumas falhas técnicas que impactam negativamente a experiência do jogador de forma implícita e explícita. A falta de sincronização entre o áudio da música e o vídeo de interpretação foi um aspecto negativo evidenciado a partir do feedback de usuários. Porém, essa falta de sincronia é natural no processo de interpretação devido à estrutura gramatical e semântico-pragmática da língua. Outro aspecto limitante identificado diz respeito à velocidade dos movimentos que precisam ser detectados pelo sistema, uma vez que movimentos muito acelerados dificultaram ou até impediram a captura de mãos durante a execução do componente.

Além disso, a análise dos resultados aponta que a diversidade de músicas permite observar diferenças de adaptação e desempenho linguístico de cada perfil de usuário. Sendo assim, para versões futuras do jogo, pode-se considerar níveis adaptativos de dificuldade ou feedback personalizado, ajustado ao perfil do jogador.

Para trabalhos futuros, pretende-se endereçar às limitações técnicas anteriormente mencionadas, a exemplo da sincronia da música e ritmo para melhorar a experiência do jogador. Outra melhoria relevante seria a otimização do sistema de detecção com a implementação de uma camada de pré-processamento de imagens, que evitaria a ocorrência de eventuais pontuações incorretas causadas por condições de ambiente nas imagens coletadas. Também se identifica como proveitoso para trabalhos futuros a utilização de modelos de Aprendizagem de Máquina para a geração de avaliações automatizadas mais precisas, bem como, o uso de dataset multi-view devido à importância de múltiplas perspectivas na captação de nuances dos sinais.

Vislumbra-se ainda como trabalhos futuros ampliar o conjunto de sinais reconhecidos pelo sistema; integrar recursos de avaliação da precisão da detecção; realizar estudos com maior número e diversidade de usuários surdos e ouvintes; investigar estratégias de feedback visual e háptico para orientar os usuários durante a sinalização; avaliar a integração do LibrasDetec a outras aplicações além do karaokê, como plataformas educacionais e tradutores automáticos.

Por fim, identifica-se que a principal contribuição deste trabalho consiste na proposição de um componente com arquitetura modular capaz de ser integrado a diferentes aplicações interativas com foco em Libras, abrindo espaço para novos estudos sobre acessibilidade digital, jogos educativos e aprendizado de línguas de sinais.

8 AGRADECIMENTOS

Agradecemos o suporte e financiamento do Ministério da Gestão e da Inovação em Serviços Públicos (MGI) do Governo Federal Brasileiro, através da Secretaria de Gestão e Inovação (SGI).

REFERENCES

[1] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (1974), 716–723. https://doi.org/10.1109/TAC.1974.1100705

[2] Diego Roberto Antunes and Janaine Daiane Rodrigues. 2021. Endless Running Game to Support Sign Language Learning by Deaf Children. In *International Conference on Human-Computer Interaction*. Springer International Publishing, Cham.

[3] G. E. P. Box and D. R. Cox. 2018. An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 26, 2 (12 2018), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

[4] Brasil. 2002. Lei nº 10.436, de 24 de abril de 2002. http://www.planalto.gov.br/ccivil_03/leis/2002/L10436.htm Acesso em: 17 jul. 2025.

[5] Alexander Brettmann, Jakob Grävinghoff, Marlene Rüschoff, and Marie Westhues. 2025. Breaking the Barriers: Video Vision Transformers for Word-Level Sign Language Recognition. *arXiv preprint arXiv:2504.07792* (2025). https://doi.org/10.48550/arXiv.2504.07792

[6] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2021), 172–186.

[7] John M Chambers. 2018. *Graphical methods for data analysis*. Chapman and Hall/CRC.

[8] William S. Cleveland. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Amer. Statist. Assoc.* 74, 368 (1979), 829–836. https://doi.org/10.1080/01621459.1979.10481038

[9] R Dennis Cook. 1977. Detection of influential observation in linear regression. *Technometrics* 19, 1 (1977), 15–18.

[10] M. Cullinan and L. L. Wood. 2024. Getting Inspired: A Qualitative Study on the Use of the Inspirisles Role-Playing Game to Teach Middle Schoolers American Sign Language. *Simulation & Gaming* 55, 2 (2024), 267–280. https://doi.org/10.1177/10468781241229614 Original work published 2024.

[11] Ralph D'Agostino and E. S. Pearson. 1973. Tests for departure from normality. Empirical results for the distributions of b2 and b1. *Biometrika* 60, 3 (12 1973), 613–622. https://doi.org/10.1093/biomet/60.3.613

[12] P. T. Dinh, T. T. Nguyen, H. Q. Le, V. H. Nguyen, T. P. Nguyen, H. N. Tran, and N. T. Nguyen. 2025. Sign Language Recognition: A Large-Scale Multi-View Dataset and Comprehensive Evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. https://openaccess.thecvf.com/content/WACV2025/papers/Dinh_Sign_Language_Recognition_A_Large-Scale_Multi-View_Dataset_and_Comprehensive_Evaluation_WACV_2025_paper.pdf

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2353–2362. https://doi.org/10.1109/ICCV.2017.256

[14] R. A. Fisher. 1992. *Statistical Methods for Research Workers*. Springer New York, New York, NY, 66–70. https://doi.org/10.1007/978-1-4612-4380-9_6

[15] Manolis Fragkiadakis and Peter van der Putten. 2021. Sign and Search: Sign Search Functionality for Sign Language Lexica. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Association for Machine Translation in the Americas, Virtual, 23–32. https://aclanthology.org/2021.mtsummit-at4ssl.3/

[16] A. Halder, A. ans Tayade. 2021. Real-time vernacular sign language recognition using mediapipe and machine learning. 2 (2021), 9–17.

[17] Valerie Henderson, Seungyon Lee, Helene Brashear, Harley Hamilton, Thad Starner, and Steven Hamilton. 2005. Development of an American Sign Language game for deaf children. Association for Computing Machinery, 70–79. https://doi.org/10.1145/1109540.1109550

[18] Carlos M. Jarque and Anil K. Bera. 1987. A Test for Normality of Observations and Regression Residuals. *International Statistical Review / Revue Internationale de Statistique* 55, 2 (1987), 163–172. http://www.jstor.org/stable/1403192

[19] Z. Li, X. Wang, Y. Wang, and J. Zhang. 2024. YOLOv5-DTW: Gesture recognition based on YOLOv5 and dynamic time warping for digital media design. *Journal of Applied Science and Engineering* 29, 2 (2024), 1–10. http://jase.tku.edu.tw/articles/jase-202602-29-02-0019

[20] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann. 2019. MediaPipe: A Framework for Perceiving and Processing Reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.

[21] C. R. Ramos. 2022. *A Língua de Sinais dos Surdos Brasileiros*. Arara Azul, Petrópolis-SP.

[22] J. B. Ramsey. 1969. Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* 31, 2 (1969), 350–371. https://doi.org/10.1111/j.2517-6161.1969.tb00796.x

[23] P. Senin. 2008. Dynamic Time Warping Algorithm Review. 855 (2008), 23–40.

[24] A. P. Shanker and A.N. Rajagopalan. 2007. Off-line signature verification using DTW. 28 (2007), 1407–1414.

[25] Luca Ulrich et al. 2024. SIGNIFY: Leveraging Machine Learning and Gesture Recognition for Sign Language Teaching through a Serious Game. *Future Internet* 16, 12 (2024), 447. https://doi.org/10.3390/fi16120447

[26] Y. Wang and X. Liu. 2019. DTWNet: A Dynamic Time Warping Network. In *Advances in Neural Information Processing Systems*, Vol. 32. http://papers.neurips.cc/paper/9338-dtwnet-a-dynamic-time-warping-network.pdf