UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Beatriz Teodoro

Desenvolvimento de ferramenta para análise de sequências de imagens e vídeos digitais em LIBRAS

São Paulo

Junho de 2012

Universidade de São Paulo

Escola de Artes, Ciências e Humanidades

Beatriz Teodoro

Desenvolvimento de ferramenta para análise de sequências de imagens e vídeos digitais em LIBRAS

Monografia apresentada à Escola de Artes, Ciências e Humanidades, da Universidade de São Paulo, como parte dos requisitos exigidos na disciplina ACH 2017 – Projeto Supervisionado ou de Graduação I, do curso de Bacharelado em Sistemas de Informação.

Modalidade:
TCC Curto (1 semestre) - individual

São Paulo

Junho de 2012

Universidade de São Paulo
Escola de Artes, Ciências e Humanidades

Beatriz Teodoro

Desenvolvimento de ferramenta para análise de sequências de imagens e vídeos digitais em LIBRAS

Orientação

_____

Prof. Dr. Luciano Antonio Digiampietri

Banca Examinadora:

_____

Profa. Dra. Ariane Machado Lima

_____

São Paulo, junho de 2012

Agradecimentos

Primeiramente gostaria de agradecer ao meu orientador Prof. Dr. Luciano Digiampietri, por sua grande dedicação e paciência durante todo o desenvolvimento do projeto, além também por sua competência, confiança, orientação e pelos conhecimentos que me proporcionou durante todas as etapas da realização deste trabalho.

Agradeço à professora de LIBRAS Maria Carolina Casati, que gravou os vídeos utilizados neste trabalho, além também de ajudar na anotação das imagens e validação do sistema.

Agradeço também a toda a minha família, em especial aos meus pais e meus irmãos, que me aturaram nos momentos de estresse e que de forma carinhosa tiveram muita paciência comigo, me apoiando sempre nos momentos de dificuldade, me dando força e coragem, além também de me proporcionarem um conforto para que eu tivesse o tempo necessário para me dedicar aos trabalhos e provas da universidade.

Por fim, agradeço a todos os meus amigos e colegas de curso, que estiveram ao meu lado durante esses anos de batalha e muito estudo, contribuindo com sugestões, críticas e bastante apoio.

Dedicatória

Gostaria de dedicar esse trabalho ao meu irmão Rafael Teodoro, o qual me influenciou positivamente na escolha da realização do curso de Sistemas de Informação e em especial na escolha da área de pesquisa a qual este trabalho faz parte.

Glossário

ACM: ACM *Digital Library.*

ANN: *Artificial Neural Network.*

CBIR: *Content-based Image Retrieval.*

CDFD: *Combined Discriminative Feature Detectors.*

DTW: *Dynamic Time Warping.*

HMM: *Hidden Markov Models.*

HVC: *Hierarchial Voting Classification.*

IEEE: IEEExplore.

LIBRAS: Língua Brasileira de Sinais.

PET-SI: Programa de Educação Tutorial de Sistemas de Informação.

RNN: *Recurrent Neural Network.*

SVM: *Support Vector Machines.*

SOFM: *Self-Organizing Feature Maps.*

Resumo

Nos últimos anos, é possível observar que o empenho em facilitar a comunicação entre surdos e pessoas que não conhecem uma língua gestual tem aumentado, mas o processo de reconhecimento e tradução de línguas de sinais ainda é pouco desenvolvido.

Este trabalho tem o intuito de apresentar um estudo e uma primeira implementação de uma ferramenta para análise de sequências de imagens segmentadas relacionadas à Língua Brasileira de Sinais (LIBRAS), com o objetivo de reconhecer sinais dinâmicos e traduzir palavras expressas nas sequências dessas imagens para o português, a fim de facilitar a comunicação entre surdos e pessoas que não tem o conhecimento de LIBRAS.

Palavras chaves: Reconhecimento de LIBRAS, tradução de LIBRAS, processamento de vídeo, processamento de imagens.

Lista de Figuras

Lista de Tabelas

Sumário

Introdução

Atualmente pode-se observar que o número de pessoas que não possuem deficiência auditiva que sabem se comunicar utilizando linguagem gestual é muito pequeno, tornando extremamente difícil a comunicação de surdos com estas pessoas. Muitos esforços são realizados para criar uma ligação entre as pessoas sem deficiência auditiva e as surdas, mas ainda há muito trabalho a ser feito. O reconhecimento e tradução automáticos de línguas gestuais ainda é uma área que está em estado inicial de desenvolvimento.

O foco desse trabalho é o reconhecimento de Língua Brasileira de Sinais (LIBRAS), com a finalidade de simplificar a comunicação entre surdos conversando em LIBRAS e ouvintes que não saibam essa língua. O reconhecimento será realizado através do processamento de imagens e vídeos digitais de pessoas se comunicando em LIBRAS, focando em sinais que utilizam apenas as mãos.

Podemos ver o processamento de um vídeo como o processamento de uma sequência de imagens onde se pode tirar proveito da ordem na qual essas imagens se encontram, com a finalidade de descrever ou reconhecer certas características do vídeo.

Esse projeto tem como objetivo desenvolver uma ferramenta para analisar sequências de imagens segmentadas relacionadas a LIBRAS, a fim de identificar sinais dinâmicos e efetuar uma tradução intermodal desses sinais para a Língua Portuguesa (SEGALA, 2010).

Esse trabalho apresenta uma abordagem extensível que só lida com a segmentação da mão e o reconhecimento de sinais a partir desta segmentação, com base num banco imagens construído especialmente para esse projeto, utilizando dois (dos cinco) parâmetros relacionados à sinalização de línguas gestuais: configuração da mão e orientação da palma.

O projeto em questão é uma parte essencial de um projeto de Iniciação Científica, que também está sendo realizado pela aluna deste trabalho de conclusão de curso, através do Programa de Educação Tutorial de Sistemas de Informação (PET-SI), intitulado "Desenvolvimento de biblioteca de funções para o processamento de imagens e vídeos digitais" (TEODORO, 2012). O objetivo geral desse projeto de iniciação é desenvolver um conjunto de funções (ou biblioteca) para o processamento de imagens e vídeos. Esse conjunto

de funções será composto por funções básicas para o processamento de imagens e também de funções específicas para o processamento de imagens relacionadas a LIBRAS.

O trabalho de iniciação científica supracitado desenvolveu ferramentas para o processamento de imagens, incluindo segmentação e extração de características e teve parte de seus resultados publicados em (DIGIAMPIETRI et al., 2012). O trabalho focou no processamento de imagens individuais relacionadas a LIBRAS. No presente trabalho de conclusão de curso, pretende-se adicionar uma funcionalidade fundamental, porém bastante complexa, para o processamento de LIBRAS: o processamento de vídeos e não apenas imagens individuais a fim de identificar sinais dinâmicos. Neste trabalho de conclusão de curso foi desenvolvida a primeira versão da ferramenta para reconhecimento de sinais dinâmicos. Durante todo o segundo semestre deste ano, esta ferramenta será estendida, verificada e validada.

Objetivos

Objetivo Geral

O objetivo geral deste projeto foi desenvolver uma ferramenta para analisar sequências de imagens já segmentadas, referentes à comunicação em LIBRAS (língua gestual-visual), com a finalidade de reconhecer sinais dinâmicos e realizar a tradução de palavras representadas pelas sequências dessas imagens para a Língua Portuguesa (língua oral-auditiva), utilizando ao menos a configuração da mão e a orientação da palma.

Objetivos Específicos

Os objetivos específicos deste trabalho foram:

- Aprimorar os mecanismos de segmentação de imagens que já haviam sido desenvolvidos durante a iniciação científica;
- Dada uma sequência de imagens segmentadas relacionadas a LIBRAS, descobrir a palavra que está sendo sinalizada.
    - Estudo de trabalhos correlatados para entender o estado da arte e as diferentes maneiras de processar sequências de imagens de línguas gestuais.
    - Especificação, implementação e teste de uma ferramenta para o reconhecimento automático de palavras em vídeos de LIBRAS.

Revisão Bibliográfica

Com o objetivo de ter um conhecimento elementar do estado da arte sobre reconhecimento de línguas de sinais, foi realizada uma revisão sobre o assunto. Para ter uma revisão precisa, foram adotadas as seguintes fontes e critérios:

- IEEE: IEEExplore.
- ACM: ACM Digital Library.

Para cada uma das fontes de dados foi utilizada a expressão *"sign language recognition"* como chave de busca. A seleção dos artigos identificados pela estratégia de busca foi feita com base na leitura do *abstract* de cada um dos resultados obtidos, levando em conta os seguintes critérios de inclusão e exclusão:

- Inclusão: Trabalhos que apresentem técnicas de reconhecimento de língua de sinais.
- Exclusão: Trabalhos que apresentem técnicas de reconhecimento de sinais utilizando luva de dados e/ou sensores.

Os dez primeiros artigos de cada base foram selecionados com base nos critérios de relevância de cada uma e nos critérios estabelecidos anteriormente. Os artigos selecionados foram lidos na íntegra e foi realizado um levantamento dos pontos mais importantes de cada um deles, identificando que método de aquisição de imagem, que parâmetros e que técnicas de reconhecimento foram utilizados.

Nas línguas sinalizadas há cinco parâmetros relacionados à realização de sinais: (a) configuração da mão (há 63 configurações diferentes); (b) posição; (c) orientação da palma da mão (KLIMA; BELLUGI, 1979); (d) movimento; e (e) expressões não manuais (por exemplo, faciais).

A Tabela 1 contém um resumo da revisão bibliográfica, com os pontos mais relevantes de todos os artigos selecionados. Dos artigos revisados, apenas no trabalho de Caridakis, Asteriadis e Karpouzis (2011) não são utilizados os parâmetros de configuração da mão e movimento, pois o foco desse trabalho é tratar a incorporação de características não manuais, como expressões faciais, o olhar e a pose da cabeça, no reconhecimento de língua de sinais. Na obtenção dos vídeos, a maioria dos trabalhos utilizou ambientes controlados, com objetivo

de diminuir a complexidade do processo de segmentação das imagens que é feito posteriormente. Mesmo utilizando o critério de exclusão citado anteriormente, quatro dos vinte trabalhos selecionados utilizaram luvas especiais, dessa forma, não foi preciso processar vídeos, mas sim os sinais feitos com estas luvas. Dos restantes, cinco apresentaram o uso de luvas coloridas, assim como é utilizado neste trabalho.

Referência	**Método de Aquisição de Imagens**	**Parâmetros Utilizados**	**Técnica de Reconhecimentos**
(PAULRAJ et al., 2011)	vídeos gravados em um estúdio, sem luvas	configuração da mão, posição e movimento	*Artificial Neural Network (ANN)*
(MAEBATAKE et al., 2008)	não consta	configuração da mão, orientação da palma, posição e movimento	*Multi-Stream HMM*
(YU et al., 2011)	vídeos gravados com luvas coloridas	configuração da mão, orientação da palma posição e movimento	*Multi-Stream HMM*
(HUANG; JIANG; ZHAO, 2010)	vídeos convencionais	configuração da mão, orientação da palma, posição, movimento e expressões não manuais	*Gabor Wavelet Transforms*
(QUAN, 2010)	vídeos de mãos com fundo branco	configuração da mão, orientação da palma e movimento	*Support Vector Machines (SVM)*
(BAUER; HIENZ, 2000)	vídeos gravados em um estúdio com luvas de cores diferentes	configuração da mão, orientação da palma, posição e movimento	*Hidden Markov Models (HMM)*
(KUMARAGE et al., 2011)	vídeos obtidos por duas câmeras em um estúdio	configuração da mão, orientação da palma, posição, movimento e expressões não manuais	Distância entre imagens
(SANDJAJA; MARCOS, 2009)	vídeos gravados com luvas multicoloridas	configuração da mão, orientação da palma, posição, movimento e expressões não manuais	*Hidden Markov Models (HMM)*
(THEODORAKIS; KATSAMANIS; MARAGOS, 2009)	vídeos gravados em estúdio, sem luvas	configuração da mão, posição e movimento	*Product HMM*

(FANG; GAO; MA, 2001)	uso de luvas de dados *(data gloves)*	configuração da mão, orientação da palma, posição e movimento	*Self-Organizing Feature Maps/ Hidden Markov Models (SOFM/HMM)*
(TEN HOLT et al., 2011)	vídeos 3D gravados em estúdio, sem luvas	configuração da mão, orientação da palma, posição e movimento	*Combined Discriminative Feature Detectors (CDFD)*
(FANG; GAO; ZHAO, 2003)	uso de luvas de dados *(data gloves)*	configuração da mão, orientação da palma, posição e movimento	*Hierarchical Decision Trees* SOFM/HMM
(ZHANG et al., 2004)	vídeos gravados em um estúdio com luvas coloridas	configuração da mão, orientação da palma, posição e movimento	*Tied-Mixture Density HMM*
(MICHAEL; METAXAS; NEIDLE, 2009)	vídeos obtidos por quatro câmeras em um estúdio, sem luvas	configuração da mão, orientação da palma, posição, movimento e expressões não manuais	*Stacked Support Vector Machine*
(ZHANG et al., 2005a)	uso de luvas de dados *(data gloves)*	configuração da mão, orientação da palma, posição e movimento	*Bosted HMM*
(CARIDAKIS et al., 2008)	vídeos gravados em um estúdio, sem luvas	configuração da mão, orientação da palma, posição e movimento	*Self-Organizing Maps*
(ZHANG et al., 2005b)	vídeos gravados em um estúdio com luvas coloridas	configuração da mão, orientação da palma, posição e movimento	*Hierarchial Voting Classification (HVC)/ Hidden Markov Models (HMM)*
(CARIDAKIS; ASTERIADIS; KARPOUZIS, 2011)	vídeos gravados em um estúdio, sem luvas	expressões não manuais	*Recurrent Neural Network (RNN)*
(JIANG; YAO, H.; YAO, G., 2004)	uso de luvas de dados *(data gloves)*	configuração da mão, posição e movimento	*Dynamic Time Warping (DTW)/ ISODATA classifier, HMM*
(DIMOV; MARINOV; ZLATEVA, 2007)	vídeos gravados em estúdio, sem luvas	configuração da mão, posição e movimento	*Content-based Image Retrieval (CBIR)*

Tabela 1: Resumo da revisão bibliográfica.

Atualmente, existem duas abordagens principais para reconhecimento de língua de sinais: a abordagem visual, onde os dados são obtidos através de uma câmera de vídeo, e a abordagem baseada em dispositivos eletromecânicos, como luvas de dados. A primeira abordagem é extremamente adequada para ser aplicada na vida diária, pois é mais conveniente para o usuário. Além disso, a comunicação através de um vídeo pode oferecer ao

deficiente auditivo uma liberdade de comunicação de longa distância. Em contrapartida exige um pré-processamento sofisticado. Na segunda abordagem existem desconfortos e limitações para o usuário utilizar uma luva de dados, porém esta abordagem facilita o reconhecimento, evitando problemas enfrentados na primeira abordagem, como o de segmentação e monitoramento das mãos.

Os principais algoritmos de reconhecimento de língua de sinais utilizados ultimamente são: *Hidden Markov Mode* (HMM), Redes Neurais Artificiais (RNA), entre outros. Paulraj et al. (2011) desenvolveram um sistema utilizando vídeos gravados por uma webcam, apresentando uma precisão de 92,58% para reconhecer gestos de 44 fonemas em inglês americano, através das RNA. *Multi-Stream* HMM é utilizado por Maebatake et al. (2008) para discutir a importância da posição e do movimento da mão no reconhecimento de sinais, assim como por Yu et al. (2011) no reconhecimento da Língua de Sinais de Taiwan, através de vídeos gravados com luvas coloridas. O algoritmo HMM é utilizado por Bauer e Hiens (2000) no reconhecimento da Língua de Sinais Alemã, através de uma câmera de vídeo de cor única com o uso de luvas coloridas. Nesse sistema uma precisão de 91,7% pode ser alcançada com base num léxico de 97 sinais.

HMM também é utilizado por Sandjaja e Marcos (2009) no reconhecimento dos números em Língua Filipina de Sinais, atingindo uma precisão média de 85,52%, utilizando vídeos gravados com o uso de luvas multicoloridas. Zhang et al. (2004) também apresentam um sistema que utiliza vídeos gravados em estúdio com o uso de luvas multicoloridas como entrada. O sistema utiliza o algoritmo *Tied-Mixture Density* HMM no reconhecimento, obtendo 92,5% de precisão, com base em 439 palavras frequentemente utilizadas na Língua Chinesa de Sinais. No sistema de reconhecimento apresentado por Zhang et al. (2005b), também foram utilizados vídeos gravados em estúdio com o uso de luvas coloridas e o reconhecimento sobre um vocabulário de 223 em Língua Chinesa de Sinais é feito através de *Hierarchial Voting Classification* (HVC) e *Continuous Hidden Markov Models* (CHMM), mostrando que a abordagem HVC baseada em conjuntos supera a abordagem convencional CHMM simples, com melhoria relativa de 30,3% da precisão.

Product HMM é utilizado por Theodorakis, Katsamanis e Maragos (2009) no reconhecimento de vídeos gravados em estúdio, expressando um vocabulário de 93 sinais em Língua de Sinais Grega. Huang, Jiang e Zhao (2010) apresentaram um estudo sobre o

reconhecimento de língua de sinais baseado em *Gabor Wavelet Transforms*, utilizando vídeos convencionais como entrada. Michael, Metaxas e Neidle (2009) utilizaram quatro câmeras sincronizadas para a captura de vídeos sem o uso de luvas e Stacked Support Vector Machine no reconhecimento de 64 sequências de vídeos de Língua de Sinais Americana. *Support Vector Machines* (SVM) é utilizado por Quan (2010) no reconhecimento da Língua Chinesa de Sinais, obtendo 95,55% como taxa média de reconhecimento, utilizando 30 grupos de imagens do alfabeto manual Chinês, obtidos através de vídeos de mãos com fundo branco. Dimov, Marinov e Zlateva (2007) utilizam Recuperação de Imagens por Conteúdo para o reconhecimento do alfabeto da Língua de Sinais Búlgara, através de videoclipes gravados em estúdio.

Abordagens instrumentais são apresentadas por Zhang et al. (2005a) e Fang, Gao e Ma (2001), utilizando *cybergloves* e rastreadores *Pohelmeis 3SPACE-posicion* como dispositivos de entrada. Zhang et al. (2005a) apresentaram que o algoritmo *Bosted* HMM melhora a precisão do reconhecimento em cerca de 3% comparado ao CHMM tradicional, utilizando um vocabulário de 102 palavras frequentemente utilizadas na Língua Chinesa de Sinais. Fang, Gao e Ma (2001) demonstraram que a combinação de *Self-Organizing Feature Maps* (SOFMs) com HMMs aumenta a precisão do reconhecimento em 5% comparado a um sistema baseado em HMM, utilizando um vocabulário de 208 palavras em Língua Chinesa de Sinais.

Esses dispositivos de entrada também foram utilizados por Jiang, Yao H. e Yao G. (2004) e Fang, Gao e Zhao (2003) em seus trabalhos. Jiang, Yao H. e Yao G. (2004) demonstraram o uso de uma arquitetura multicamadas no reconhecimento da Língua Chinesa de Sinais, combinado o algoritmo *Dynamic Time Warping* (DTW)/ISODATA com HMM. Fang, Gao e Zhao (2003) desenvolveram um sistema que utiliza *Hierarchical Decision Trees e* SOFM/HMM no reconhecimento, utilizando um grande vocabulário de 5.113 palavras em Língua Chinesa de Sinais. O método proposto nesse sistema reduz surpreendentemente o tempo de reconhecimento em 11 vezes e também melhora a taxa de reconhecimento em cerca de 0,95% em relação ao SOFM/HMM simples.

Pelos artigos analisados, nota-se um grande volume de trabalhos que utilizam a abordagem visual para o reconhecimento de língua de sinais, assim como foi utilizado nesse

trabalho, evitando as limitações que a abordagem baseada em dispositivos eletromecânicos oferece.

Algo que chamou a atenção na revisão foi a utilização de luvas coloridas e/ou de ambientes controlados na gravação dos vídeos na maioria dos trabalhos, as mesmas ferramentas utilizadas na gravação dos vídeos utilizados nesse trabalho, facilitando o pré-processamento realizado posteriormente, como o processo de segmentação das imagens.

Por fim, outro ponto que chamou a atenção foi a marcante presença do algoritmo HMM no reconhecimento de língua de sinais na maioria dos artigos revisados. Nesse trabalho, foi utilizado o algoritmo de distância de Levenshtein (*Levenshtein Distance*) na implementação da ferramenta para o reconhecimento de língua de sinais, diferente das ferramentas encontradas na literatura, definida posteriormente na Subseção 4.2.

Metodologia

A metodologia deste trabalho consistiu, primeiramente, do estudo de diversos trabalhos relacionados ao reconhecimento de línguas de sinais publicados nos últimos anos, através de uma revisão bibliográfica, a qual foi detalhada na Seção 3, a fim de obter uma concepção das técnicas de reconhecimento mais utilizadas.

A partir do estudo realizado, foi especificada e implementada uma ferramenta para o reconhecimento automático de algumas palavras dentro de vídeos de pessoas se comunicando em LIBRAS, tratada na Subseção 4.3. A implementação dessa ferramenta foi feita na linguagem JAVA[1]. A implementação resultante foi testada em diferentes conjuntos de imagens apresentados na Subseção 4.2, segmentadas através da técnica apresentada na Subseção 4.1. A Subseção 4.4 apresenta a estratégia de teste utilizada e os resultados da execução dos testes foram analisados e validados com a ajuda de uma especialista do domínio dessas imagens, apresentados na seção de resultados.

[1] http://www.java.com

Segmentação das imagens

A segmentação das imagens nesse trabalho foram feitas utilizando uma extensão de uma função implementada em (DIGIAMPIETRI et al., 2012), que é uma técnica semi-automática (supervisionada) para segmentar utilizando como referência um conjunto de imagens segmentadas manualmente. A função segmenta uma imagem em seis regiões: os cinco dedos (mínimo, anelar, médio, indicador e polegar) e a palma.

A extensão implementada neste TCC dividiu o processo de segmentação em 3 etapas. A versão original utilizava apenas uma etapa de processamento para identificação das 6 regiões da mão, mas foi constatado que este processo gera muitos ruídos quando os vídeos incluem a pessoa (e não apenas a mão) e alguns objetos no ambiente.

Devido a maior quantidade de objetos nos vídeos utilizados neste TCC, a primeira etapa do novo algoritmo de segmentação consiste em identificar a região em que a mão com a luva multicolorida encontra-se em cada frame do vídeo e gerar novas imagens delimitadas por essa região. As etapas subsequentes de processamento consistem da segmentação propriamente dita. O processamento é semiautomático, no qual um usuário precisa segmentar manualmente algumas imagens de vídeo e estas imagens segmentadas são usadas como referência para um algoritmo de classificação automática utilizando *RotationForest*. Com esse processo foi obtida uma taxa de acertos da segmentação de 93,07%, em um teste realizado com 1400 pixeis de 26 imagens segmentadas manualmente (considerando a segmentação pixel a pixel), utilizando a técnica de validação *10-fold cross-validation* (DIGIAMPIETRI et al., 2012).

Na extensão do algoritmo desenvolvida para este TCC, esta classificação semiautomática está dividida em duas etapas: separação do fundo e segmentação das partes da mão. A divisão em duas etapas da classificação semiautomática melhorou essa taxa de acertos da segmentação de 93,07% para 95,06%.

A Figura 1 apresenta um exemplo da realização do pré-processamento definido acima. A imagem a esquerda da figura apresenta um frame do vídeo utilizado. A imagem do centro apresenta a imagem gerada após o rastreamento da mão com a luva. Por fim, a imagem da direita apresenta a imagem gerada após a segmentação.

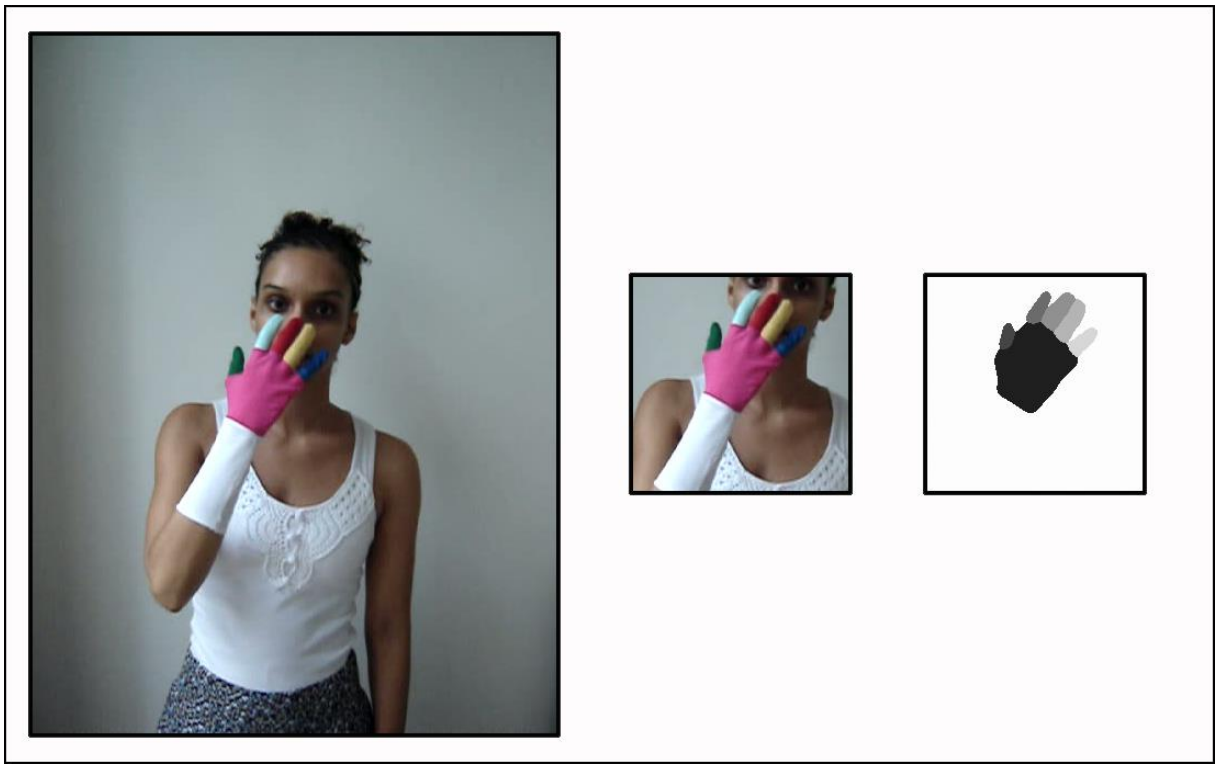

Figura 1: Exemplo de resultado após o pré-processamento.

Dos cinco vídeos gravados para este TCC, primeiramente utilizamos o processo de segmentação semi-automática detalhado acima, utilizando como base apenas imagens geradas por dois deles. Isto foi feito para verificar a influência de problemas na segmentação nos resultados produzidos pelo algoritmo de reconhecimento de LIBRAS.

Após termos usado a floresta de rotação gerada para os dois vídeos, que tiveram uma segmentação manual de algumas de suas imagens, para segmentar os demais vídeos, constatamos muitos ruídos nas imagens resultantes após a segmentação. Os problemas ocorridos na segmentação foram causados por a especialista ter usado roupas diferentes e a luminosidade ter sido diferente na gravação dos vídeos. Assim, foi constatado que o processo de segmentação é dependente da etapa manual (segmentação manual de algumas imagens) e por isso a segmentação manual de imagens foi feita para todos os vídeos (de 10 a 30 imagens por vídeo). Neste trabalho não foram explorados qual seria a quantidade mínima necessária de imagens a serem segmentadas por vídeo para garantir uma segmentação de qualidade e nem

mesmo a influência da quantidade de imagens segmentadas manualmente no resultado na segmentação.

Após ter sido feito essas adaptações a taxa de acerto da ferramenta implementada aumentou significativamente, conforme será visto na Seção 5.

Banco de imagens

Os conjuntos de imagens utilizados foram gerados a partir de cinco vídeos de uma especialista expressando sinais em LIBRAS e um vídeo adicional apenas sinalizando as 63 configurações de mão, utilizando uma luva multicolorida na mão direita. A gravação dos vídeos foi feita utilizando uma câmera fotográfica digital, operando no modo de configuração automático. Cada vídeo possui 25 frames por segundo e cada frame do vídeo apresenta resolução de 480x640 pixels.

A especialista sinalizou, em dias diferentes e utilizando roupas distintas, as 63 diferentes configurações de mão existentes em um dos vídeos e nos vídeos restantes sinalizou um conjunto de 39 palavras básicas (casa, cachorro, noite, dia, etc.) escolhidas por ela. Essas palavras escolhidas foram sinalizadas uma vez em cada vídeo. As imagens geradas foram pré-processadas utilizando a ferramenta apresentada na Subseção 4.1.

As imagens extraídas destes vídeos formaram o banco de imagens utilizado na verificação da ferramenta implementada, apresentada na Subseção 4.3. As imagens geradas pelo vídeo onde são sinalizadas as 63 diferentes configurações de mão existentes foram utilizadas para gerar o conjunto de imagens segmentadas utilizado no processo de reconhecimento da configuração da mão. As imagens geradas pelos demais vídeos, onde é sinalizado um conjunto de 39 palavras básicas, após sofrerem o processo de segmentação, foram utilizadas para formar os conjuntos de imagens utilizados nos testes e validação da ferramenta, contendo as 39 palavras sinalizadas em cada.

Durante a iniciação científica e as etapas de revisão bibliográfica deste TCC, foram procurados bancos de imagens públicos, contendo imagens segmentadas relacionadas a

línguas de sinais. Infelizmente nenhum banco público foi encontrado e por isso a necessidade da produção de um banco de imagens neste TCC. A falta de bancos/bases de dados de referência para línguas de sinais ou mesmo de ferramentas que facilitem a criação e disponibilização destes bancos tem motivado trabalhos recentes para o oferecimento deste tipo de recurso (WAGNER et al., 2012).

Especificação e implementação da ferramenta

Para a implementação da ferramenta, foram utilizadas *strings* para representar as sequências de imagens já segmentadas que expressam as palavras em LIBRAS, sendo cada caractere das strings a simbolização da configuração em que a mão se encontra em cada uma das imagens da sequência (conforme apresentado, há 63 configurações diferentes). A Tabela 2 contém as configurações de mão existentes e os caracteres utilizados.

 1: **0**	 2: **1**	 3: **2**	 4: **3**	 5: **4**	 6: **5**	 7: **6**
 8: **7**	 9: **8**	 10: **9**	 11: **:**	 12: **;**	 13: <	 14: =
 15: >	 16: **?**	 17: **@**	 18: **A**	 19: **B**	 20: **C**	 21: **D**
 22: **E**	 23: **F**	 24: **G**	 25: **H**	 26: **I**	 27: **J**	 28: **K**
						

29: **L**	30: **M**	31: **N**	32: **O**	33: **P**	34: **Q**	35: **R**
 36: **S**	 37: **T**	 38: **U**	 39: **V**	 40: **W**	 41: **X**	 42: **Y**
 43: **Z**	 44: **[**	 45: \	 46: **]**	 47: ^	 48: _	 49: `
 50: **a**	 51: **b**	 52: **c**	 53: **d**	 54: **e**	 55: **f**	 56: **g**
 57: **h**	 58: **i**	 59: **j**	 60: **k**	 61: **l**	 62: **m**	 63: **n**

Tabela 2: Configurações de mão e caracteres utilizados.

Cada palavra sinalizada corresponde a um conjunto de imagens (frames do vídeo). A identificação do início e fim da sinalização de uma palavra foi feita manualmente. Cada imagen de cada palavra foi convertida para um caractere que corresponde à configuração de mão em que a mão estava naquela imagem. Os caracteres de cada palavra sinalizada foram concatenados em uma string. A Figura 2 contém um exemplo de sequências de imagens sendo simbolizadas através de caracteres. Nesse exemplo as sequências de imagens representam a palavra “bom” em dois vídeos diferentes.

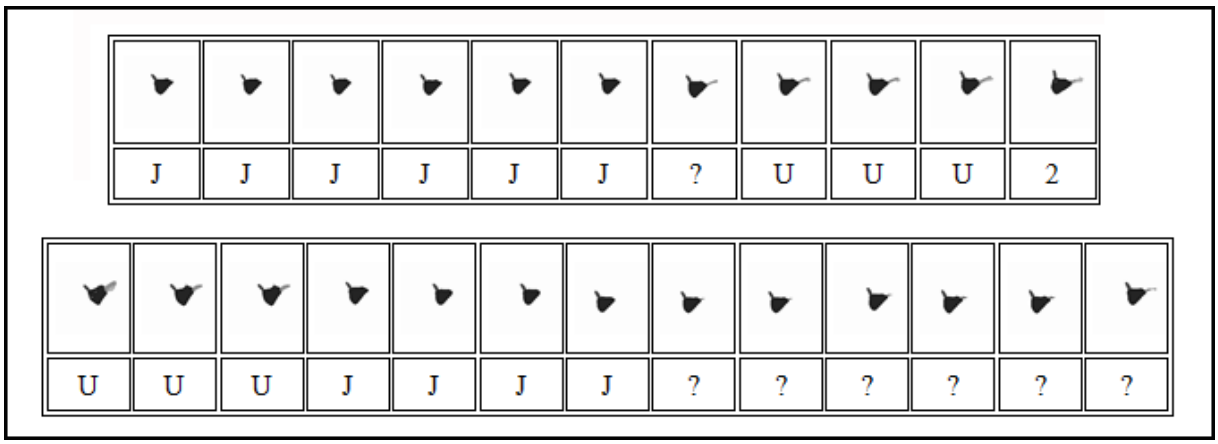

Figura 2: Simbolização da palavra "bom" em dois vídeos diferentes.

A identificação da configuração de mão correspondente a cada imagem foi feita de maneira automática, utilizando um descritor de imagem desenvolvido especialmente para imagens relacionadas a línguas de sinais (Digiampietri et. Al (2012)), nas quais as mãos estão segmentadas em 6 partes (palma da mão e cada um dos dedos). A identificação da orientação da palma não foi realizada por já estar implícita na configuração da mão.

O descritor utiliza um vetor de características que contém as propriedades extraídas das imagens. Para o descritor em questão, foram utilizados dois extratores de características: área proporcional de cada segmento e o de posição relativa de cada segmento. O primeiro extrator calcula o tamanho proporcional de cada um dos segmentos da mão em relação ao todo. Desta forma, este extrator descreve a imagem em 6 valores reais. O segundo extrator calcula a posição relativa de cada segmento em relação ao centro de gravidade da mão. Sendo assim, este extrator caracteriza cada imagem em 12 valores (6 coordenadas bidimensionais).

Com base nesse vetor de características, são calculadas as similaridades entre as imagens através de uma função de distância. Nesse trabalho as similaridades entre as imagens foram mensuradas através da distância euclidiana. Considerando os pontos A = (a_1, a_2, ... , a_n) e B = (b_1, b_2, ... ,b_n), a distância euclidiana entre esses pontos num espaço n-dimensional é calculada da seguinte maneira:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Em suma, o reconhecimento da configuração da mão de cada imagem é realizado através da medida da distância entre as características das imagens já classificadas com as características das novas imagens, de acordo com o descritor apresentado.

A Figura 3 apresenta uma ferramenta que foi desenvolvida para comparar as imagens de um diretório com uma imagem de consulta. As imagens que se encontram mais próximas ao centro da tela são as imagens mais próximas da imagem de consulta. A imagem mais próxima da imagem de consulta, ou seja, a imagem mais semelhante à imagem de consulta classifica essa imagem.

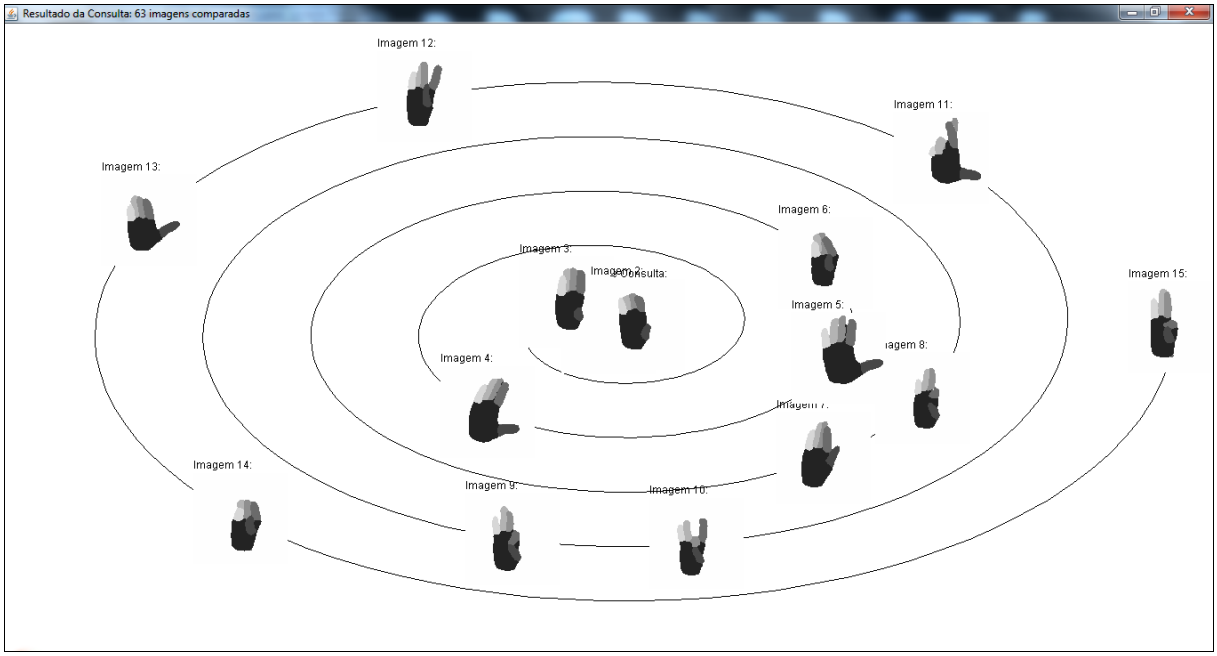

Figura 3: Resultado de busca utilizando o descritor de imagem apresentado.

Para identificar a similaridade entre as palavras, ou seja, entre as strings, foi implementado o algoritmo de distância de Levenshtein (*Levenshtein Distance*), popularmente conhecido como algoritmo para calcular a *distância de edição*. Essa técnica avalia a similaridade entre duas strings baseando-se no número mínimo de operações necessárias para transformar uma string em outra. As possíveis operações são: inserção, exclusão e substituição. Uma matriz é montada a partir do tamanho das strings, onde são configurados os custos de cada operação. Ao final das comparações entre os caracteres das strings, a distância é dada pela última posição da matriz, sendo a distância zero a indicação de que as strings são idênticas (LEVENSHTEIN, 1966).

A função de análise da sequência das imagens foi implementada utilizando esse algoritmo para comparar a string de uma palavra de entrada (devidamente codificada) com as strings que representam as 39 palavras já armazenadas no banco de dados. Os resultados obtidos com a execução dessa função são apresentados adiante, na Seção 5.

Estratégia de teste

Para a realização dos testes, através da execução da ferramenta para o reconhecimento automático de palavras sinalizadas em LIBRAS, tratada na Subseção 4.3, foram utilizados grupos de 39 palavras comuns escolhidas por uma especialista. Cada palavra foi gravada 5 vezes (em dias e condições diferente, conforme apresentado).

Pelo fato da técnica desenvolvida utilizar distância para comparar as palavras sinalizadas, optou-se por utilizar a técnica de validação *10-fold cross-validation*, onde cada acerto significa: dada a uma nova sinalização de uma palavra (ou seja, sinalização do conjunto de testes) ela estiver mais próxima (de acordo com o critério de distância apresentado) de uma outra sinalização da mesma palavra (pertencente ao conjunto de treinamento).

Resultados

Em um teste inicial realizado utilizando três grupos com 39 palavras (cada grupo extraído de um vídeo diferente, onde são expressas as mesmas palavras), foi obtida uma taxa de acerto de 34,38% para o reconhecimento destas palavras, comparando-se todas as palavras (strings) contra todas, através da ferramenta utilizada. Em um classificador aleatório a chance de acertar o reconhecimento de uma palavra codificada seria apenas 1/39, ou seja, a taxa de acerto seria de 2,56%. Essa taxa de acerto de 34,38% foi obtida utilizando imagens que foram segmentadas utilizando o pré-processamento detalhado anteriormente na Subseção 4.1. Assim como é apresentado na Subseção 4.1, o modo como havia sido feita a segmentação primeiramente gerava imagens com grandes erros de segmentação, que influenciaram negativamente na taxa de acerto obtida. Quando desconsideradas essas palavras, a taxa de acerto subiu para 66,66%. Para a obtenção dessa taxa foram excluídas quase metade das palavras utilizadas anteriormente (57 das 3*39). O critério utilizado para fazer essa exclusão foi selecionar todas as palavras que apresentassem mais de 1/3 das suas imagens com problema de segmentação. Os problemas considerados foram: a não localização da mão ou a junção da mão com outro objeto (rosto, calça, saia, etc).

Com esses resultados constatamos que a ferramenta de reconhecimento de sinais implementada é muito sensível a qualidade da segmentação. Após terem sido feitas algumas modificações no processo de segmentação das imagens, detalhadas na Subseção 4.1, foi realizado um novo teste usando agora cinco grupos com 39 palavras utilizados anteriormente. Com a realização desse novo teste obtivemos uma taxa de acerto de 52.22%.

A Figura 4 contém alguns exemplos interessantes de strings semelhantes geradas para as mesmas palavras. Esses exemplos são positivos, pois mesmo tendo sido gerado alguns símbolos (letras) diferentes para configurações de mão muito semelhantes dentro das sequências de imagens, que representam as palavras apresentadas, a função implementada foi capaz de reconhecer ambas as sequências de entrada como sendo a mesma palavra.

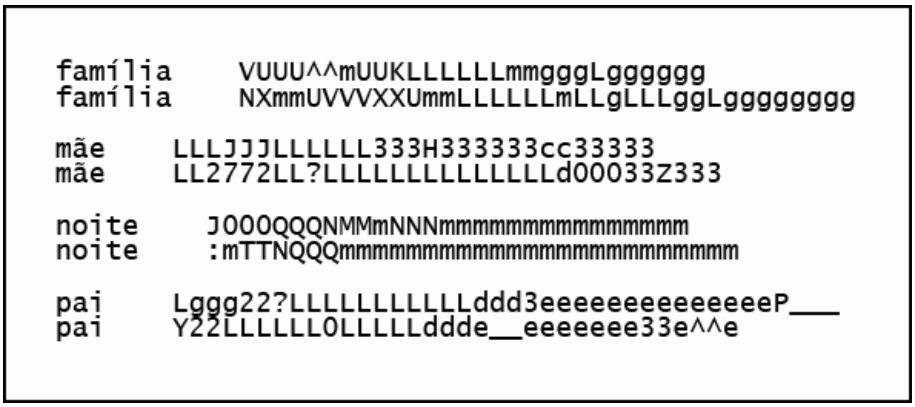

Figura 4: Exemplos de strings semelhantes geradas para as mesmas palavras.

Discussão

Pode-se observar que, devido principalmente a problemas encontrados na segmentação das imagens, os resultados gerados através da execução da ferramenta implementada foram satisfatórios. Mesmo com a melhoria da taxa de acertos da segmentação através da utilização da classificação semiautomática dividida em duas partes, ainda será necessário que seja realizado um aprimoramento ainda maior da função de segmentação, a fim de gerar melhores resultados.

Essa influência negativa na taxa de acerto por parte da segmentação pode ser observado através da mudança da taxa de acerto de 34,38%, após terem sido feitas algumas modificações no processo de segmentação realizado, um novo teste gerou uma taxa de 52,22%. Desconsiderando as palavras com grandes erros de segmentação essa taxa passou para 66,66%. Ao se testar apenas as palavras que não possuíam nenhum erro na segmentação, esta taxa de acerto ultrapassou 85%, porém restaram tão poucas palavras que esta taxa nem foi apresentada ou discutida neste trabalho.

Através do aprimoramento do segmentador, e da um aperfeiçoamento da ferramenta implementada, será possível aumentar significativamente a taxa de acerto.

Vale ainda observar que a técnica desenvolvida é baseada em distância, assim, um conjunto maior de dados (com mais sinalizações da mesma palavra) a tendência é obtermos melhores resultados. Além disso, um conjunto maior de sinalizações da mesma palavra permite que outras técnicas sejam incorporadas na classificação de palavras, por exemplo, o uso de K-Vizinhos (atualmente estamos usando apenas o vizinho mais próximo na classificação).

Uma segunda lição aprendida é sobre a segmentação. A técnica desenvolvida é eficiente quando é feita a segmentação manual de um número suficiente de imagens do vídeo a ser segmentado. Assim, é importante como complementação deste trabalho analisar a influência do número de imagens segmentadas na qualidade final da segmentação.

Conclusão

O objetivo geral deste trabalho foi cumprido com a construção de uma ferramenta que reconhece sinais dinâmicos e os traduz para a Língua Portuguesa, através da análise de sequências de imagens que representam palavras sinalizadas em LIBRAS. Através dos resultados obtidos, pode se concluir que a ferramenta implementada é bastante sensível à qualidade da segmentação das imagens.

Para a continuidade deste trabalho, que será feita no projeto de Iniciação Científica ao qual esse trabalho faz parte (TEODORO, 2012), o próximo passo será montar um novo banco de imagens através da realização de experimentos mais controlados, a fim de melhorar o processo de segmentação. Com o aprimoramento da técnica de segmentação utilizada, tornando a mais eficiente e eficaz, será possível obter resultados mais expressivos no processo de tradução das palavras expressadas em LIBRAS através da ferramenta implementada nesse trabalho.

Para aprimoramento do reconhecimento de LIBRAS propriamente dito, pretende-se utilizar um algoritmo de alinhamento local ao invés do algoritmo de distância de edição. Este tipo de algoritmo é bastante utilizado em bioinformática para a identificação de sobreposições entre sequências de nucleotídeos ou aminoácidos e potencialmente terá bons resultados para o reconhecimento de LIBRAS. Além disso, este tipo de ferramenta permite que a execução seja feita sobre todas as imagens do vídeo, sem a necessidade da identificação manual dos frames (ou imagem) nos quais cada palavra começa ou termina.

Com trabalho futuro também pretende-se estender o trabalho desenvolvido, tratando dos cinco parâmetros relacionados à sinalização de línguas gestuais, a fim de se obter um sistema mais robusto.

Referências Bibliográficas

BAUER, B. e HIENZ, H.. **Relevant Features for Video-Based Continuous Sign Language Recognition.** In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000., p. 440–445.

CARIDAKIS, G.; ASTERIADIS, S.; KARPOUZIS, K.. **Non-manual cues in automatic sign language recognition**. In Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '11). ACM, New York, NY, USA, 2011. Article 43, 4 pages. DOI: 10.1145/2141622.2141673.

CARIDAKIS, G.; DIAMANTI, O.; KARPOUZIS, K.; MARAGOS, P.. **Automatic sign language recognition: vision based feature extraction and probabilistic recognition scheme from multiple cues**. In Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '08), Fillia Makedon, Lynne Baillie, Grammati Pantziou, and Ilias Maglogiannis (Eds.). ACM, New York, NY, USA, 2008. Article 89, 8 pages. DOI: 10.1145/1389586.1389687.

DIGIAMPIETRI, L.; TEODORO, B.; SANTIAGO, C.; OLIVEIRA, G.; ARAÚJO, J. **Um sistema de informação extensível para o reconhecimento automático de LIBRAS**. SBSI 2012 - Trilhas Técnicas (*Technical Tracks*), 2012.

DIMOV, D.; MARINOV, A.; ZLATEVA, N.. **CBIR approach to the recognition of a sign language alphabet**. In Proceedings of the 2007 international conference on Computer systems and technologies (CompSysTech '07), Boris Rachev, Angel Smrikarov, and Dimo Dimov (Eds.). ACM, New York, NY, USA, 2009. Article 96 , 9 pages. DOI: 10.1145/1330598.1330700.

FANG, G.; GAO, W.; MA, J.. **Signer-Independent Sign Language Recognition Based on SOFM/MMM.** In Proceedings of the IEEE ICCV Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, Vancouver, British Columbia, Canada, 2001, p. 90-95.

FANG, G.; GAO, W.; ZHAO, D.. **Large Vocabulary Sign Language Recognition Based on Hierarchical Decision Trees.** In Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI ‘03). ACM, New York, NY, USA, 2003, p. 125–131. DOI: 10.1145/958432.958458.

HUANG, Z.; JIANG, D.; e ZHAO, W.. **Study of Sign Language Recognition Based on Gabor Wavelet Transforms.** In Proceedings of the 2010 International Conference on Computer Design and Applications (ICCDA), Qinhuangdao, Hebei, China, 2010. Vol. 1, p. V1–151 –V1–154.

JIANG, F.; YAO, H.; YAO, G.. **Multilayer architecture in sign language recognition system**. In Proceedings of the 6th international conference on Multimodal interfaces (ICMI '04). ACM, New York, NY, USA, 352-353. DOI: 10.1145/1027933.1028010.

KLIMA, E.; BELLUGI, U.. **The signs of language**. Cambridge University Press, 1979.

KUMARAGE, D.; FERNANDO, S.; FERNANDO, P.; MADUSHANKA, D.; SAMARASINGHE, R.. **Real-time Sign Language Gesture Recognition Using Still-Image Comparison & Motion Recognition.** Proceedings of the 2011 6th IEEE International Conference on Industrial and Information Systems (ICIIS), Kandy, Sri Lanka, 2011, p. 169–174.

LEVENSHTEIN, V. I. **Binary codes capable of correcting deletions, insertions and reversals**. *Soviet Physics Doklady*, 1966, *10*, 707-710 .

MAEBATAKE, M.; SUZUKI, I.; NISHIDA, M.; HORIUCHI, Y.; KUROIWA, S.. **Sign Language Recognition Based on Position and Movement Using Multi-Stream HMM.** In Proceedings of the 2008 Second International Symposium on Universal Communication, ISUC '08, Washington, DC, USA, 2008, p. 478–481.

MICHAEL, N.; METAXAS, D.; NEIDLE, C**.. Spatial and Temporal Pyramids for Grammatical Expression Recognition of American Sign Language.** In Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS). ACM, New York, NY, USA, 2009, p. 75–82. DOI: 10.1145/1639642.1639657.

PAULRAJ, M.P.; YAACOB, S.; AZALAN, M.S.Z.; PALANIAPPAN, R.. **A Phoneme Based Sign Language Recognition System using 2D Moment Invariant Interleaving feature and Neural Network.** In Proceedings of the 2011 IEEE Student Conference on Research and Development (SCOReD), Cyberjaya, Selangor, Malasy, 2011, p. 111–116.

QUAN, Y.. **Chinese Sign Language Recognition Based On Video Sequence Appearance Modeling.** In Proceedings of the 2010 the 5th IEEE Conference on Industrial Electronics and Applications (ICIEA), Taichung, Taiwan, 2010, p. 1537–1542.

SANDJAJA, I.N.; MARCOS, N. **Sign Language Number Recognition.** In Proceedings of the 5th International Joint Conference on INC, IMS and IDC, Seoul, Korea, 2009, p. 1503 - 1508.

SEGALA, R. R. . **Tradução Intermodal e Intersimiótica/Interlingual**: Português brasileiro escrito para Língua Brasileira de Sinais. (2010). 74 f. Dissertação (Mestrado em Estudos da Tradução) - Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis, 2010.

TEN HOLT, G. A.; VAN DOORN, A. J.; REINDERS, M. J. T.; HENDRIKS, E. A.; DE RIDDER, H.. **Human-Inspired Search for Redundancy in Automatic Sign Language Recognition**. ACM Transaction on Applied Perception (TAP). 8, 2, Article 15 (January 2011), 15 pages. DOI: 10.1145/1870076.1870083.

TEODORO, B.. **Desenvolvimento de biblioteca de funções para o processamento de imagens e vídeos digitais.** Trabalho de Iniciação Científica, EACH-USP, 2012.

THEODORAKIS, S., KATSAMANIS, A., e MARAGOS, P.. **Product-HMMs for Automatic Sign Language Recognition.** In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 2009, p. 1601 –1604.

WAGNER, P.; BORGES, G. MADEO, R.; PERES, S. **Uma Ferramenta para Construção de Conjuntos de Dados de Referência para Sistemas de Análise de Gestos Baseados em Imagens**. SBSI 2012 - Trilha Especial – Aplicativos em SI, 2012.

YU, S.-H.; HUANG, C.-L.; HSU, S.-C.; LIN, H.-W.; WANG, H.-W. **Vision-Based Continuous Sign Language Recognition using Product HMM.** In Preoceedings of the 2011 First Asian Conference on Pattern Recognition (ACPR), Pequim, China, 2011, p. 510–514.

ZHANG, L.-G.; CHEN, Y.; FANG, G.; CHEN, X.; GAO, W.. **A Vision-Based Sign Language Recognition System Using Tied-Mixture Density HMM.** In Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI ‘04). ACM, New York, NY, USA, 2004, p. 198–204. DOI: 10.1145/1027933.1027967.

ZHANG, L.-G.; CHEN, X.; WANG, C.; CHEN, Y.; GAO, W.. **Recognition of Sign Language Subwords Based on Boosted Hidden Markov Models.** In Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI ’05). ACM, New York, NY, USA, 2005a, p. 282–287. DOI: 10.1145/1088463.1088511.

ZHANG, L.-G.; CHEN, X.; WANG, C.; GAO, W.. **Hierarchical voting classification scheme for improving visual sign language recognition**. In Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05). ACM, New York, NY, USA, 2005b, p. 339-342. DOI: 10.1145/1101149.1101220.