Conjunto de Treinamento para Algoritmos de Reconhecimento de LIBRAS

Jéssica Barbosa Dias, Kleber Padovani de Souza e Hemerson Pistori

[1]Grupo de Pesquisa em Engenharia e Computação, GPEC
Universidade Católica Dom Bosco, UCDB
Av. Tamandaré, 6000, Jardim Seminário, 79117-900 Campo Grande, MS
djessy@gmail.com, kleber.padovani@terra.com.br, pistori@ucdb.br

Resumo. *Neste artigo descrevemos um banco de imagens que será utilizado em um sistema reconhecedor de gestos da Língua Brasileira de Sinais empregando modelos de Markov ocultos. Durante a análise foram extraídas, manualmente, 11 características, de aproximadamente 60.000 imagens contidas em 600 amostras de vídeos. Os resultados da análise de uma parcela dos vídeos dos gestos capturados serão utilizados durante a fase de treinamento para a alimentação das componentes dos modelos projetados. Posteriormente, a outra parcela de vídeos será utilizada nos experimentos para avaliar a eficácia de classificação dos gestos.*

1. Introdução

A interação homem-máquina guiada por sinais visuais é um rico paradigma de comunicação que amplia o domínio de aplicações computacionais e ameniza a dificuldade de interação entre os computadores e os seres humanos que possuem necessidades especiais, como os surdos. Ferramentas com interação guiada por sinais visuais trazem embutidas consigo o poder de abrigar uma parcela da sociedade que atualmente encontra-se impedida, ou com dificuldades, de utilizar os computadores através das formas habituais de comunicação convencionadas pelo homem, como o mouse e o teclado.

Porém, por se tratar de uma tema novo, a utilização deste paradigma ainda não é muito difundida dentro da comunidade de desenvolvedores de aplicações computacionais. Um dos obstáculos para a propagação da aplicabilidade desta interação é a complexidade computacional de implementação destas interfaces, que, em contrapartida, é uma tarefa trivial quando se utiliza meios tradicionais. Através da análise e reconhecimento dos gestos, posturas e expressões humanas por meio da visão computacional é possível enriquecer este paradigma de comunicação. Para isto, em determinadas aplicações de reconhecimento de gestos, a criação de um banco de imagens contribui muito no resultado final destes sistemas, podendo ser utilizado em várias fases do processo, como nas etapas de treinamento e testes de métodos baseados em aprendizagem supervisionada, entre outras.

Não foi encontrado na bibliografia pesquisada nenhum outro banco de imagens, com a mesma finalidade, para Língua Brasileira de Sinais. Em 2000 foi desenvolvido um dicionário *on-line* de Língua de Sinais, mantido pelo *Center for Sutton Movement Writing*, que possui um banco de sinais que são apresentados em *SignWriting*. Atualmente existem outros específicos, como o Auslan SignBank, que possui um banco de sinais da Língua Australiana de Sinais, e o Dicionário Digital da Língua Brasileira de Sinais, que permite ao usuário localizar textualmente palavras na língua portuguesa e visualizar o gesto correspondente em LIBRAS. Porém, este último não tem como objetivo criar uma base para treinamento e/ou testes em sistemas de reconhecimento de padrões, servindo apenas como uma relação entre palavras da língua portuguesa com os gestos da língua de sinais.

A finalidade deste artigo é descrever a criação e análise de um banco de imagens de gestos da Língua Brasileira de Sinais, a LIBRAS, que será utilizado no desenvolvimento de um sistema de reconhecimento de padrões através dos Modelos de Markov Ocultos (HMM) [2, 7]. Este sistema utiliza uma padronização descrita por Fernando Capovilla [1], que especifica de maneira detalhada os gestos contidos na Língua Brasileira de Sinais.

Este artigo é organizado em 5 seções. A seção 2 faz uma breve descrição sobre a Língua de Sinais, em particular a Língua Brasileira de Sinais; a seção 3 descreve as características pertinentes ao banco de

imagens criado neste trabalho, citando a metodologia utilizada para a aquisição e análise dos vídeos; e as seções 4 e 5 exibem o estado atual do projeto e as futuras tarefas a serem realizadas para sua conclusão.

2. Língua de Sinais

As línguas de sinais são linguagens naturais utilizadas por comunidades de pessoas que têm a oportunidade de se comunicar regularmente umas com as outras, geralmente surdas e/ou mudas, por todo o mundo, cujas formas consistem de seqüências de movimentos e configurações executados pelas partes do corpo, como mãos, braços e face, e por sinais realizados por expressões faciais, movimentos da cabeça, do torso e por posturas do corpo.

Assim como as linguagens faladas, existem várias línguas de sinais diferentes, como a Língua Brasileira de Sinais e a Língua Americana de Sinais, mas todas têm poder de expressão equivalente às linguagens faladas [6], porém, explorando meios físicos diferentes. Os gestos utilizados pelas línguas de sinais são geralmente considerados os mais estruturados, dentre várias categorias de gestos [3]. A Língua Brasileira de Sinais possui um conjunto de 46 diferentes configurações de mão [4]. Neste conjunto estão contidos os gestos referentes às letras do alfabeto, sendo que 20 destas são representadas por posturas e 6 por gestos. Utilizaremos em nosso banco os gestos da LIBRAS, dando enfoque no dialeto utilizado no estado de Mato Grosso do Sul.

3. Banco de Imagems

A Língua Brasileira de Sinais possui uma extensa variedade de gestos utilizados na formação de suas sentenças. Dentre eles, foram selecionados 50 gestos provenientes do dicionário trilíngüe [1] para compor o banco de imagens. Para executar a seleção, foram utilizados como critérios a complexidade do movimento do gesto, o nível de confusão que ele pode gerar com outros gestos já escolhidos devido à semelhança entre eles, e a exploração das diferentes configurações possíveis dos gestos.

Foram escolhidos gestos com movimentos complexos com intuito de testar a eficiência da aplicação da técnica, que será utilizada em ambientes em que o usuário movimenta seus braços, mãos e cabeça; o segundo critério foi utilizado para analisar a atuação do classificador diante de gestos semelhantes; e o terceiro para explorar de maneira enfática algumas configurações de postura do corpo, como a inflação das bochechas, por exemplo.

Como dito anteriormente, nas línguas de sinais os gestos são formados por expressões faciais, movimentos de ombros, braços, mãos, cabeça, etc. Porém, no banco de imagens criado, a classificação dos gestos não abrangerá todas estas características corporais para a composição do gesto, mas apenas os movimentos das mãos, braços e cabeça[1].

Foi utilizada uma câmera digital para obtenção das imagens, sendo que esta se manteve fixada em frente ao usuário e imóvel durante as gravações. Alguns gestos foram armazenados isoladamente e outros em conjunto. Os gestos armazenados em conjunto foram separados em arquivos distintos através do Cinelerra, um programa livre e gratuito de edição de vídeos para sistemas operacionais livres. Os arquivos de vídeo estão armazenados na estrutura de diretório convencional do sistema operacional Linux em formato MPEG-4.

Primeiramente, a captura das imagens foi realizada em um ambiente com fundo estático e uniforme, ou seja, não existem objetos atrás do usuário e o único objeto com movimento é ele próprio. Em seguida, foram coletadas imagens com fundos dinâmicos e não-uniforme em ambientes diversificados.

Foram capturadas imagens de 4 usuários, dois homens e duas mulheres, sendo que 3 são destros e 1 é canhoto. Cada usuário executou três vezes todos os gestos selecionados, produzindo 12 amostras para cada gesto, gerando um total de 600 amostras. Durante as gravações não foi utilizado nenhum recurso de apoio para o rastreamento das mãos e da face, como luvas de dados ou câmeras de infra-vermelho.

Uma postura é uma configuração estática, sem movimento, enquanto o gesto é dinâmico, ou seja, possui movimento. Por exemplo, a foto de uma mão e a filmagem de uma cabeça se deslocando da esquerda para

[1] As características de movimentação de cabeça incluem apenas a situação das bochechas, excluindo as expressões faciais e a posição da mesma.

a direita são exemplos de postura e gesto, respectivamente. De uma postura o usuário transita para outra postura, que, conseqüentemente, produz um gesto.

O reconhecimento dos gestos será realizado utilizando HMM [8, 3]. Para cada gesto foi desenvolvido um HMM e para a construção desses modelos foram escolhidas as posturas que constituem os gestos, sendo que cada uma destas posturas se relaciona diretamente a um estado do modelo. As posturas foram denominadas como constituintes do gesto de acordo com uma análise empírica visual, ou seja, foram selecionadas as posturas que são visualmente mais destacáveis nos gestos, e que a transição por elas produza o gesto. Analogamente, estas posturas podem ser comparadas aos fonemas das linguagens faladas, de forma que o uso seqüencial dos fonemas forma o som da palavra.

Por exemplo, para executar o gesto referente a “derrame cerebral”, a mão direita deve estar em *M* [1], com a palma virada para a esquerda e com os dedos apontando para cima, tocando o lado direito da testa. Em seguida, devem-se cruzar os braços em frente à cabeça, com as duas mãos fechadas e com as palmas para trás. Por fim, as mãos devem ser movidas para os lados opostos com força. Neste exemplo, as posturas consideradas como constituintes são as ilustradas pela Figura 1, conseqüentemente, o modelo construído para a representação desta expressão conterá três estados.

(a) Mão direita em *M*, com a palma virada para a esquerda e com os dedos apontando para cima, tocando o lado direito da testa;

(b) Braços cruzados em frente à cabeça, com as duas mãos fechadas e com as palmas para trás;

(c) Braços paralelos ao lado do corpo, com as duas mãos fechadas e com as palmas para trás.

Figura 1. *Posturas do gesto referente à expressão “derrame cerebral” em LIBRAS.*

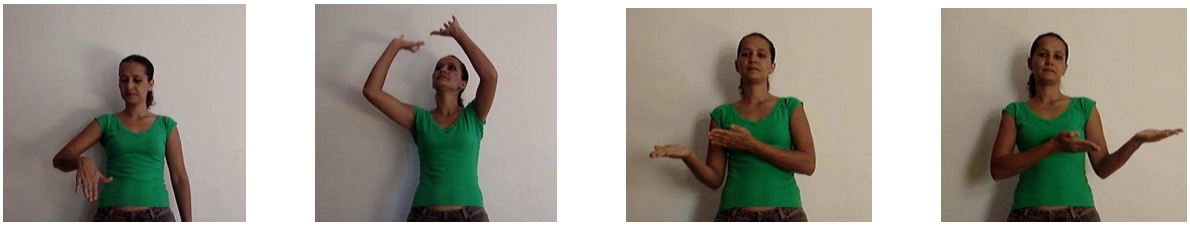
Como apenas algumas posturas têm um respectivo estado no modelo, algumas posturas são “ignoradas”, predominando apenas as que se encaixam em algum estado existente. As posturas contidas no gesto que são classificadas como desconhecidas para o modelo são aproximadas pelo sistema a uma das posturas selecionadas, obedecendo a determinado critério de classificação, e, com isto, obtemos as transições de estados de cada HMM.

Primeiramente, definimos as características que, juntas, nos auxiliam a discriminar as posturas executadas pelos usuários. As características foram escolhidas com base nas descrições e padronizações contidas em [1] e na análise dos gestos escolhidos, observando quais combinações discriminavam as posturas selecionadas de maneira única.

A primeira característica é a *posição espacial vertical*, PEV, que representa a altura da posição das mãos em relação ao corpo do usuário. Para esta característica foram estabelecidas nove possíveis variações. Duas variações da posição espacial vertical podem ser observadas nas Figuras 2(a) e 2(b), pois na primeira a mão direita do usuário está localizada entre a cintura e o peito, enquanto que na segunda está acima da cabeça.

A segunda característica é a *posição espacial horizontal*, PEH, que é a localização das mãos em relação a um eixo imaginário vertical que corta o centro do corpo do usuário. Existem 8 variações possíveis para a PEH. Na Figura 2 podem ser visualizadas duas variações desta característica, em que as mãos estão primeiramente ao lado direito do corpo, como mostra a Figura 2(c), e, em seguida, partem para o lado esquerdo, de acordo com a Figura 2(d).

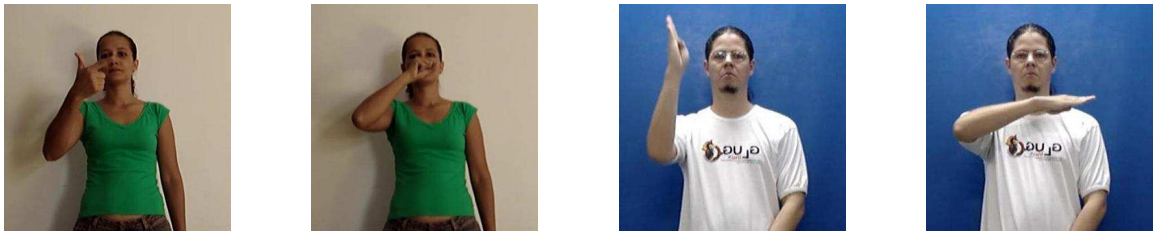
Ao executar os gestos em LIBRAS, as mãos, quando integrantes do gesto, assumem determinadas



(a) (b) (c) (d)

Figura 2. *2(a) e 2(b) Primeira e segunda posturas do gesto referente à palavra "basquete" em LIBRAS, respectivamente. As figuras 2(c) e 2(d) referem-se à primeira e à segunda postura do gesto referente à palavra "bem-vindo" em LIBRAS, nesta ordem.*

configurações, que são estabelecidas pelos dedos. Esta *configuração da mão*, CON, é a terceira característica selecionada, e pode assumir os valores contidos em um conjunto de 22 variações. Uma variação da configuração das mãos (CON) pode ser observada na Figura 3. Na Figura 3(a) o usuário está com a mão na configuração da letra "L", enquanto que em outro instante do gesto, representado pela Figura 3(b), a mão está na configuração da letra "S"[2].



(a) (b) (c) (d)

Figura 3. *3(a) e 3(b) Primeira e segunda posturas do gesto referente à palavra "inodoro" em LIBRAS, respectivamente. 3(c) e 3(d) Primeira e segunda posturas do gesto referente à palavra "mau" em LIBRAS, nesta ordem.*

A quarta e a quinta característica são a *orientação*, ORI, e a *direção*, DIP, da palma da mão. A ORI determina se a mão está em posição vertical ou horizontal, porém, em relação a um eixo imaginário que segue do pulso até a ponta do dedo médio, enquanto que a DIP, como o nome sugere, define a direção da palma em relação ao corpo do usuário[3]. Duas entre as três variações da ORI e duas entre as 7 variações da DIP podem ser observadas na Figura 3, referente ao sinal "mau" em LIBRAS, sendo que na Figura 3(c) a mão direita está na posição vertical, com a palma para a esquerda e na Figura 3(d) na posição horizontal com a palma para baixo. E, por fim, a *situação das bochechas*, SIB, indica se as bochechas do usuário encontram-se infladas ou não[4]. A SIB pode ser observada pelas Figuras 4(a) e 4(b), em que na segunda o usuário está com as bochechas infladas, o que não ocorre na primeira.

Neste sentido, cada postura do gesto é representada com a combinação destas características, sendo que todas as que se referem a informações da mão (PEV, PEH, CON, ORI e DIP) se transformam em duas características: mão direita e mão esquerda. Além disto, existe a possibilidade de uma das mãos, ou ambas, não estar presente nas posturas, o que implica na inclusão da variação *desconhecida* em cada uma destas características. Portanto, cada postura do gesto se distingüe pelo conjunto de valores das 11 características. Porém, este conjunto apresenta algumas limitações, caso o conjunto de gestos seja estendido, pois foram extraídas as características necessárias para distingir apenas os gestos selecionados. Portanto, características como expressão facial e configuração dos ombros não foram consideradas como atributos relevantes para o reconhecimento de cada gesto.

Como estamos utilizando modelos discretos, as características acima mencionadas devem receber apenas

[2]A letra "L" em libras se caracteriza pela mão fechada e os dedos indicador e polegar distendidos, e a letra "S" pela mão fechada, com o polegar cobrindo os outros dedos.

[3]As variações *para baixo* e *para cima* da DIP não são configuradas em relação ao corpo do usuário, mas em relação à mão em questão do usuário.

[4]A situação das bochechas se configura como *inflada* se uma ou ambas as bochechas estiverem infladas.

(a) (b)

Figura 4. *Duas posturas do gesto referente à palavra "demolir" em LIBRAS.*

valores discretos, desta forma, em nosso caso todos os valores possíveis para cada característica são previamente conhecidos. Com base nesta padronização os vídeos foram classificados um a um e quadro a quadro. Como um vídeo é composto de uma seqüência de imagens (quadros), foram extraídas as características acima para todas as imagens de cada vídeo, e, posteriormente, foram atribuídos valores a elas, através de análise visual.

Em cada vídeo de gesto coletado existem aproximadamente 100 quadros. Ou seja, foram analisados 600 vídeos de 100 quadros, extraindo 11 características em cada quadro, totalizando em média 660.000 características. Em cada quadro foram observadas as mãos e face do usuário, extraindo as posições vertical e horizontal de cada mão, as orientações e direções de cada palma, a configuração executada por cada mão e a situação das bochechas do usuário. As informações oriundas desta análise serão úteis para a formação das matrizes de probabilidades (matrizes de transição de estados e geração de símbolos) de cada HMM. As descrições dos gestos foram retiradas do dicionário LIBRAS [1], porém, alguns gestos foram modificados, conseqüentemente, suas descrições também foram alteradas em nosso documento. Como exemplo, o gesto referente à palavra "babá" no dicionário inicia com a mão direita em 1, palma para trás, tocando a ponta do indicador abaixo do olho direito. Porém, esta movimentação inicial não existe no dialeto LIBRAS de Mato Grosso do Sul, portanto, alteramos sua descrição.

4. Experimentos a Serem Realizados

Estão sendo desenvolvidas 4 aplicações que processarão as imagens contidas no banco de imagens e produzirão os resultados do sistema: um gerador de vídeos segmentados, um extrator de características, um gerador de HMM e um classificador de vídeos. Todas as aplicações utilizarão recursos existentes em uma plataforma de apoio ao desenvolvimento de aplicações guiadas por sinais visuais, denominada SIGUS[5], que oferece suporte a este tipo de aplicação.

De maneira resumida, o gerador de vídeos segmentados receberá como entrada um arquivo de vídeo original e trará como resultado um novo arquivo de vídeo segmentado utilizando um método de segmentação específico. O extrator de características receberá estes vídeos segmentados e fará a extração dos atributos, que serão processados, produzindo os respectivos HMMs para cada gesto. As características extraídas pelo extrator são de natureza analógica, entretanto, os modelos utilizados são de natureza discreta. Por este motivo, antes da construção dos modelos, os valores provenientes da extração devem ser discretizados. Por fim, o classificador receberá como entrada vídeos originais e fará sua classificação, baseada nos modelos criados anteriormente.

Como exemplo de utilização destas aplicações, considere que $\frac{2}{3}$ dos vídeos dos gestos referentes às palavras pertencentes ao banco de imagens serão utilizados para treinamento e que $\frac{1}{3}$ será utilizado para a classificação. Primeiramente os vídeos de treinamento serão enviados ao segmentador e submetidos ao processo de segmentação, de forma que permaneçam nos vídeos apenas os objetos de interesse, que no caso são as mãos e a face, excluindo o resto.

Subseqüentemente, os vídeos segmentados de cada palavra são utilizados como entrada pelo extrator, a fim de extrair os atributos que resumem as distribuições de *pixels* resultantes para auxiliar no processo de

[5]Maiores informações sobre a plataforma SIGUS podem ser encontradas no site www.gpec.ucdb.br/sigus.

discriminação, como o total de *pixels*, centro de massa e desvios padrões das mãos em cada quadro de cada vídeo. Os valores extraídos passam pela discretização e são utilizados para a construção do modelo referente a cada palavra.

Neste momento todos os modelos de cada palavra estão criados e devidamente ajustados. Com isto, é possível realizar a classificação dos vídeos restantes. Aleatoriamente se escolhe um vídeo referente a alguma palavra. Este vídeo é submetido ao processo de segmentação e extração de atributos, porém sem a construção de um novo modelo. Os dados obtidos da extração são enviados ao módulo de classificação que informa qual o modelo que tem a maior probabilidade de ter gerado esta amostra e classifica a entrada como a palavra referente a este modelo.

5. Considerações Finais

Neste trabalho investimos na utilização de uma técnica bastante utilizada no reconhecimento de fala e muito bem conceituada, porém, aplicada ao reconhecimento de gestos: os HMMs [5, 9]. A popularização dos modelos de Markov ocultos tem origem no sucesso de sua aplicação no reconhecimento de fala. Nós, utilizamos-no para definir gestos em um sistema reconhecedor, em que cada HMM corresponde a um gesto reconhecido pelo sistema.

Primeiramente, serão tratados apenas os vídeos com fundos estáticos e uniformes, e, posteriormente, vídeos com fundos mais complexos. O sistema computacional reconhecedor será derivado da aplicação LIBRAS, contida na plataforma SIGUS, incluindo as melhorias necessárias para seu funcionamento. É necessário ainda analisar e escolher os algoritmos de segmentação, bem como os métodos e atributos para a extração das características das imagens.

Atualmente, os HMMs construídos possuem apenas os estados, o conjunto de símbolos e as transições, faltando ainda as probabilidades de transição de estados, geração de símbolos e de ocorrência inicial. Após a etapa de classificação manual dos gestos de treinamento serão alcançados estes componentes restantes dos modelos e a partir deste momento começarão a ser realizados os experimentos e análise de eficiência do sistema.

6. Agradecimentos

Este trabalho recebe apoio financeiro da Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul, FUNDECT, da Financiadora de Estudos e Projetos, FINEP, e do Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq.

Referências

[1] Fernando C. Capovilla and Walkiria D. Raphael. *Dicionário Enciclopédico Ilustrado Trilíngüe da Língua de Sinais Brasileira*, volume I e II. São Paulo, SP: Edusp, Imprensa Oficial, 2002.

[2] M. Morita and L. S. Oliveira. Introdução aos modelos escondidos de markov. Technical report, PPGIA-PUCPR, Curitiba-Brazil, November 1998.

[3] Sylvie C. W. Ong and Surendra Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):873–891, 2005.

[4] H. Pistori. *Tecnologia Adaptativa em Engenharia de Computação: Estado da Arte e Aplicações*. PhD thesis, Universidade de São Paulo, São Paulo, São Paulo, Brasil, 2003.

[5] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE Computer Graphics and Applications*, 77(2):42–53, 1989.

[6] Wendy Sandler and Diane Lillo-Martin. *Natural Sign Languages*, pages 533–562. Oxford: Blackwell, 2001.

[7] Thad Starner. Visual recognition of american sign language using hidden markov models. Technical Report Master's Thesis, MIT, Feb 1995, Program in Media Arts & Sciences, MIT Media Laboratory, 94.

[8] Donald Tanguay. Hidden markov models for gesture recognition. Technical report, Cambridge: Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science., 93.

[9] Andrew D. Wilson and Aaron F. Bobick. Hidden markov models for modeling and recognizing gesture under variation. *IJPRAI*, 15(1):123–160, 2001.