Estudo do estado da arte das técnicas de reconhecimento das línguas de sinais por computador

J.P. de Santiago Neto e L.S. Oquendo

Resumo — **No Brasil existem milhões de pessoas portadoras de deficiências auditivas. Para parte destas pessoas, a língua principal utilizada para comunicação é a língua de sinais LIBRAS (Língua Brasileira de Sinais) e não o português. Uma contribuição importante na inclusão digital dessas pessoas seria o reconhecimento de LIBRAS pelos computadores. Porém, o reconhecimento das línguas de sinais pelos computadores ainda é uma área aberta à pesquisa.**
Nestes últimos anos houve vários avanços no âmbito computacional. Tendo em conta estes avanços, este estudo expõe o estado da arte, no mundo, das técnicas de reconhecimento das línguas de sinais por computador.

Palavras-chave **— Visão Computacional, Reconhecimento de Línguas de Sinais, LIBRAS.**

I. Introdução

Segundo o IBGE, existem no Brasil cerca de seis milhões de pessoas portadoras de deficiências auditivas. Para parte destas pessoas, a língua principal utilizada para comunicação é a língua de sinais LIBRAS (Língua Brasileira de Sinais) e não o português. As línguas de sinais não são meras transcrições das línguas faladas, sendo compreensível a dificuldade de comunicação existente entre pessoas não surdas e surdas, mesmo que por meio da escrita, uma vez que pessoas surdas podem não ser alfabetizadas em português.

Uma contribuição importante na inclusão digital destas pessoas seria o reconhecimento de LIBRAS pelos computadores. Porém, o reconhecimento das línguas de sinais pelos computadores ainda é uma área aberta à pesquisa.

Duas são as abordagens usadas para o reconhecimento das línguas de sinais. Uma faz uso de luvas especiais para facilitar a detecção e o rastreamento dos movimentos das mãos [1]. A outra não faz uso de luvas nem de qualquer material [1], se baseando nas técnicas de Visão Computacional. Ela facilita o acesso dos usuários à tecnologia por não exigir gastos adicionais com luvas especiais. Visão Computacional é a abordagem que atualmente recebe a maior atenção dos pesquisadores e é o foco deste estudo.

J. P. de Santiago Neto, Centro Universitário Estácio do Ceará (Estácio FIC), jnetinho90@gmail.com
L. S. Oquendo, Centro Universitário Estácio do Ceará (Estácio FIC), luciano.oquendo@gmail.com

A grande dificuldade no reconhecimento das línguas de sinais pelos computadores está na complexidade do problema. A interpretação de frases em uma língua de sinais exige desde técnicas de processamento de imagens até técnicas avançadas de inteligência artificial.

Nestes últimos anos houve vários avanços no âmbito computacional. Tendo em vista estes avanços, este estudo realizou uma pesquisa para levantar o estado da arte das técnicas de reconhecimento pelos computadores das línguas de sinais, no mundo, e LIBRAS no Brasil.

II. Complexidade do problema de reconhecimento das línguas de sinais pelos computadores

A complexidade do problema de reconhecimento por computador dos gestos nas línguas de sinais vem das seguintes características das línguas de sinais:

- Os gestos nas línguas de sinais não se limitam apenas a movimentos das mãos e braços, os chamados gestos manuais, mas também de sinais realizados por expressões faciais, movimentos da cabeça, do dorso e por posturas do corpo. Desta maneira, para o reconhecimento dos gestos nas línguas de sinais é necessário uma observação simultânea dessas partes do corpo.
- A quantidade de sinais é muito grande, da ordem de milhares, alguns difíceis de diferenciar de outros.
- Um mesmo gesto nas línguas de sinais é feito diferentemente por pessoas diferentes e até pela mesma pessoa quando repetido.
- Uma frase nas línguas de sinais é uma seqüência contínua de gestos, sendo difícil a detecção de onde termina um gesto e começa o seguinte.

Tendo em vista estas características, para o reconhecimento automático por computador de gestos nas línguas de sinais se busca soluções que ofereçam precisão (para minimizar erros), velocidade de processamento (para suportar aplicações em tempo real) e facilidade de escalabilidade (para suportar os milhares de sinais).

III. Técnicas para o reconhecimento das línguas de sinais

A. Visão computacional

A Visão Computacional procura integrar as áreas de processamento digital de imagens e de inteligência

artificial, tendo como objetivo a obtenção de algoritmos capazes de interpretar o conteúdo visual de imagens.

Algumas das funcionalidades comuns na maioria dos sistemas de Visão Computacional são:

- Aquisição de imagem: é o primeiro passo para um sistema de Visão Computacional. Trata-se do processo de aquisição de imagens a partir de câmeras de vídeo.
- Pré-processamento: processo que visa facilitar a identificação de objetos (face e mãos, por exemplo) que compõem uma imagem.
- Segmentação: processo realizado para isolar regiões de pontos da imagem pertencentes a objetos (face e mãos, por exemplo) para a extração de características.
- Extração de características: extração de características matemáticas (tais como formato, localização, movimento) dos objetos que compõem uma imagem.
- Classificação: processamento de alto nível que classifica os objetos através da comparação das características dos objetos segmentados com aquelas de objetos de classes previamente estabelecidas.

B. Técnicas de detecção da face e das mãos

Dentre as técnicas utilizadas para a detecção das partes do corpo envolvidas na representação dos gestos, destacam-se as técnicas baseadas na cor da pele [2][3] e nas formas das partes do corpo rastreadas [4].

As técnicas baseadas na cor da pele têm a vantagem de serem mais simples e mais rápidas de processar computacionalmente, mas têm a desvantagem de serem mais sujeitas a erros de detecção por causa de variações de iluminação do ambiente. Segundo Nascimento [5], a detecção de cor da pele no espectro visível pode ser uma tarefa bastante desafiadora, visto que a cor da pele em uma imagem é sensível a vários fatores. Nascimento [5] apresenta uma comparação de três tipos diferentes de clusters, seguidos de operações morfológicas para detecção de pele humana em imagens digitais. Os modelos de cor utilizados foram: RGB, YCbCr e HSV. A comparação visa apontar qual modelo é mais adequado para a detecção automática da pele.

As técnicas baseadas nas formas das partes do corpo têm a vantagem de serem mais imunes a variações de iluminação do ambiente, mas têm a desvantagem de um processamento mais lento. Dentre as técnicas se destaca aquela baseada no algoritmo *AdaBoost* [4] [11] [13] usando o método de Viola e Jones [6] [7]. Com o avanço da velocidade de processamento dos processadores, demonstrações apontam que esta é uma técnica promissora para o reconhecimento em tempo real de gestos nas línguas de sinais [8].

C. Técnicas para tratar dos movimentos das mãos

É necessário detectar os movimentos das mãos para reconhecer cada gesto das línguas de sinais. São duas as técnicas para detectar os movimentos das mãos.

A primeira técnica se baseia no rastreamento dos movimentos das mãos [8]. É uma técnica mais complexa, pois exige o rastreamento preciso dos movimentos das mãos na seqüência de imagens.

A segunda técnica se baseia na aparência dos movimentos [8]. É uma técnica mais simples, pois não exige o rastreamento preciso dos movimentos das mãos. Ela é baseada na localização das mãos nas poses consideradas marcantes para o reconhecimento dos gestos das línguas de sinais.

D. Técnicas para tratar as variações espaciais e temporais dos gestos

Classificar os gestos das línguas de sinais tem ainda uma complexidade adicional, pois um mesmo gesto é feito diferentemente de pessoa para pessoa, ou até pela mesma pessoa quando repetido. Isso faz com que um mesmo gesto tenha diferenças no espaço e no tempo. Ainda, é necessário identificar quando termina um gesto e começa outro, através da seqüência de imagens [10]. Entre as técnicas que são usadas como solução para tratar as variações espaciais e temporais destacam-se aquelas baseadas nos modelos de Cadeia de Markov de 1ª ordem (*1st order Markov chain*) [4] [8], HMM (*Hidden Markov Model*) [9] e DTW (*Dynamic Time Warping*) [10]. Implementar o modelo da Cadeia de Markov de 1ª ordem é mais simples que o modelo HMM. As cadeias de Markov e HMM modelam processos estocásticos, sendo baseados em estatística. Há, porém, várias restrições ao modelo DTW [14]. Para superar estas restrições, Lichtenauer [12] propõe o uso do modelo SDTW (*Statistical Dynamic Time Warping*) no reconhecimento das línguas de sinais.

E. Técnicas para permitir a escalabilidade

Para superar o problema de escalabilidade, foram propostos modelos de representação dos gestos onde cada gesto é dividido em fonemas [4] [8]. Os fonemas são unidades visuais que podem ser reutilizados em diferentes palavras (como na língua falada), tornando a escalabilidade do sistema de reconhecimento menos complexa.

Duas são as abordagens para definir os fonemas. Uma abordagem onde os fonemas são definidos tendo por base a lingüística das línguas de sinais [4] [8]. Outra abordagem, que não se baseia na lingüística, onde os fonemas são definidos com o propósito de aumentar a eficiência computacional [10].

Kadir [4] apresenta um sistema monocular (por exemplo, uma câmara WEB) capaz de reconhecer gestos das línguas de sinais em muito maior número do que outras abordagens. Baseado em fonemas, o sistema apresenta quatro elementos-chave: (i) A detecção da cabeça e das mãos é feita com base em *boosting*, o que elimina a necessidade da sensível segmentação pela cor da pele. (ii) A descrição das atividades centradas no corpo, o que supera os problemas de posicionamento da câmera, de calibragem e do utilizador. (iii) Uma classificação em dois estágios, em que um primeiro estágio gera um nível de descrição lingüística de alto nível da atividade que, naturalmente, generaliza e, conseqüentemente, reduz o treinamento. (iv) Um segundo estágio composto por um banco de classificadores que não necessita de HMM, reduzindo ainda mais os requisitos de treinamento. O resultado é um sistema capaz de funcionar em tempo real, gerando altos índices de reconhecimento para grandes léxicos com uma única instância de treino por sinal. Foram obtidas taxas de classificação tão alta como 92% para um léxico de 164 palavras com exigências extremamente baixas de treinamento, superando outras abordagens onde milhares de exemplos de treinamento são necessários.

Cooper [8] discute também o reconhecimento de língua de sinais tendo por base os fonemas visuais. Apresenta três tipos de fonemas visuais para análise: sinais aprendidos baseados na aparência dos movimentos, sinais aprendidos baseados no rastreamento dos movimentos das mãos em 2D e sinais aprendidos baseados no rastreamento dos movimentos das mãos em 3D usando o *Kinect*. Os fonemas visuais são então combinados em um classificador que faz o reconhecimento do sinal. Duas opções de classificador são apresentadas. O primeiro utiliza *Markov Models* para codificar as mudanças temporais entre fonemas visuais (intra e inter fonemas). O segundo faz uso do *Sequential Pattern Boosting* para selecionar as características discriminativas, ao mesmo tempo em que codifica a informação temporal. A abordagem *Sequential Pattern Boosting* se revelou mais robusta a ruído e obteve um desempenho de 76% nos testes independentes do utilizador, onde as Cadeias de Markov alcançaram somente 54%.

Dias e Souza [15] implementaram um sistema capaz de reconhecer sinais em LIBRAS, integrado à plataforma SIGUS de apoio ao desenvolvimento de aplicações guiadas por sinais visuais. O reconhecimento dos gestos é realizado utilizando o modelo HMM. 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 posturas que são visualmente mais marcantes nos gestos, e que a transição por elas produza o gesto. Estas posturas constituem os fonemas.

Como apenas as posturas denominadas como marcantes têm um respectivo estado no modelo HMM, algumas posturas são "ignoradas", predominando apenas as mais importantes. 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 a direita são exemplos de postura e gesto, respectivamente. De uma postura o usuário transita para outra postura, que, conseqüentemente, produz um gesto, e, com isto, obtêm-se as transições de estados do modelo. No entanto, um modelo de Markov oculto não se constitui apenas de estados e transições, mas também necessita das matrizes de probabilidades de transição de estados e de geração de símbolos, além do conjunto de probabilidades iniciais.

Foram definidas as características que, juntas, auxiliam a discriminar os gestos executados pelos usuários. As características foram escolhidas com base nas descrições e padronizações contidas no Dicionário Enciclopédico Ilustrado Trilíngue da Língua de Sinais Brasileira, e na análise dos gestos escolhidos, observando quais combinações discriminavam os gestos selecionados de maneira única. Estas características são as seguintes: i) posição espacial vertical da mão; ii) posição espacial horizontal da mão; iii) configuração da mão; iv) orientação da mão; v) direção da palma da mão; e vi) situação das bochechas.

IV. Conclusões

As técnicas de reconhecimento por computador das línguas de sinais em geral, e LIBRAS em particular, tiveram uma substancial evolução. Visão computacional se estabeleceu como a abordagem mais utilizada, por ser mais acessível aos usuários. Os algoritmos de detecção de imagens, tal como o *Adaboost* de Viola e Jones, são mais robustos às variações do ambiente, prescinde de ambientes controlados e é de fácil uso pelos usuários em geral. O reconhecimento de língua de sinais baseado em fonemas visuais tornou o mecanismo de reconhecimento mais modular e viabilizou o reconhecimento de grande número de símbolos, gerando a perspectiva de vir a atender às necessidades reais das comunicações das línguas de sinais.

Referências

[1] Shujjat Khan, Gourab Sen Gupta, Donald Bailey, Serge Demidenko, Chris Messom, "Sign Language Analysis and Recognition: A Preliminary Investigation", 24th International Conference Image and Vision Computing New Zealand (IVCNZ 2009).

[2] Alex T. S. Carneiro, Paulo C. Cortez, Rodrigo C. S. Costa, "Reconhecimento de Gestos da LIBRAS com Classificadores Neurais a partir dos Momentos Invariantes de Hu", Anais da Interaction South America 09, 27 a 29 de Novembro de 2009, São Paulo.

[3] Feng-Sheng Chen, Chih-Ming Fu, Chung-Lin Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models", Image and Vision Computing 21 (2003) 745–758.

[4] Timor Kadir, Richard Bowden, Eng Jon Ong and Andrew Zisserman, "Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition", British Machine Vision Conference, 2004.

[5] Andréia V. Nascimento, Michelle M. Mendonça, Juliana G. Denipote, Maria Stela V. Paiva, "Comparação de Clusters para Detecção da Pele", Biblioteca Digital Brasileira de Computação, disponível em http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Trabalho?id=14657, acesso em junho de 2013.

[6] Paul Viola and Michael J. Jones, "Robust Real-Time Object Detection", Cambridge Research Laboratory, Technical Report Series, fevereiro de 2001, disponível em http://www.hpl.hp.com/techreports/Compaq-DEC/CRL-2001-1.pdf, acesso em junho de 2003.

[7] Paul Viola and Michael J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2001.

[8] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, Richard Bowden, "Sign Language Recognition using Sub-Units", Journal of Machine Learning Research 13 (2012) 2205-2231, disponível em http://jmlr.org/papers/volume13/cooper12a/cooper12a.pdf, acesso em junho 2013.

[9] Richard Bowden, David Windridge, Timor Kadir, Andrew Zisserman, and Michael Brady, "A Linguistic Feature Vector for the Visual Interpretation of Sign Language", European Conference on Computer Vision, Springer-Verlag, 2004.

[10] Han, Junwei; Awad, George; Sutherland, Alistair, "Subunit Boundary Detection for Sign Language Recognition Using Spatio-temporal Modelling", The 5th International Conference on Computer Vision Systems, 2007.

[11] Eng-Jon Ong and Richard Bowden, "A boosted classifier tree for hand shape detection", Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition, pages 889-894, 2004.

[12] Lichtenauer, J.F., Hendriks, E.A. and Reinders, M.J.T. "Sign Language Recognition by Combining Statistical DTW and Independent Classification", Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:30 , Issue: 11), Nov. 2008, pages 2040 – 2046.

[13] Hardy Francke, Javier Ruiz-del-Solar and Rodrigo Verschae, "Real-time Hand Gesture Detection and Recognition using Boosted Classifiers and Active Learning", Second Pacific Rim Symposium, PSIVT 2007 Santiago, Chile, December 17-19, 2007 Proceedings, pp 533-547.

[14] Eamonn J. Keogh and Michael J. Pazzani, "Derivative Dynamic Time Warping", In First SIAM International Conference on Data Mining (SDM'2001), disponível em http://www.ics.uci.edu/~pazzani/Publications/keogh-kdd.pdf, acesso em junho de 2013.

[15] Jéssica Barbosa Dias e Kleber Padovani de Souza, "Reconhecimento de Gestos Utilizando Modelos de Markov Ocultos", Universidade Católica Dom Bosco (UCDB), monografia, novembro de 2006, disponível em http://www.yumpu.com/pt/document/view/12552536/monografia-gpec-ucdb, acesso em junho de 2013.